Journal of Machine Learning Research 21 (2020) 1-36 Submitted 12/18; Revised 4/20; Published 9/20

Conic Optimization for Quadratic Regression Under Sparse Noise

Igor Molybog IGORMOLYBOG@BERKELEY.EDU
Department of Industrial Engineering and Operations Research
University of California
Berkeley, CA 94720, USA

Ramtin Madani RAMTIN.MADANI@UTA.EDU
Department of Electrical Engineering
University of Texas
Arlington, TA 76102, USA

Javad Lavaei LAVAEI@BERKELEY.EDU
Department of Industrial Engineering and Operations Research
University of California
Berkeley, CA 94720, USA

Editor: Massimiliano Pontil

Abstract

This paper is concerned with the quadratic regression problem, where the goal is to find the unknown state (numerical parameters) of a system modeled by a set of equations that are quadratic in the state. We focus on the setting when a subset of equations of fixed cardinality is subject to errors of arbitrary magnitudes (potentially adversarial). We develop two methods to address this problem, which are both based on conic optimization and are able to accept any available prior knowledge on the solution as an input. We derive sufficient conditions for guaranteeing the correct recovery of the unknown state for each method and show that one method provides a better accuracy while the other one scales better to large-scale systems. The obtained conditions consist in bounds on the number of bad measurements each method can tolerate without producing a nonzero estimation error. In the case when no prior knowledge is available, we develop an iterative-based conic optimization technique. It is proved that the proposed methods allow up to half of the total number of measurements to be grossly erroneous.The efficacy of the developed methods is demonstrated in different case studies, including data analytics for a European power grid.

Keywords: nonlinear regression, conic programming, bad data detection

1. Introduction

Nonlinear regression aims to find the parameters of a given model based on observational data. One may assume the existence of a potentially nonlinear continuous function $f(\mathbf{x};\mathbf{a})$ defined over the set of all possible models $\mathbf{x} \in \mathcal{X}$ and all possible inputs $\mathbf{a} \in \mathcal{A}$, where the goal is to estimate the true model given a set of imperfect measurements y_i's:

$$y_i = f(\mathbf{x};\mathbf{a}_i) + \eta_i, \qquad \forall i \in \{1,...,m\} \tag{1}$$

©2020 Molybog Igor, Ramtin Madani and Javad Lavaei.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v21/18-881.html.

In this formulation, the unknown error vector $\boldsymbol{\eta}$ could be the measurement noise with modest values. However, a more drastic scenario corresponds to the case where the vector $\boldsymbol{\eta}$ is sparse and its nonzero entries are allowed to be arbitrarily large. Under this circumstance, *a priori* information about the probability distribution of the sparse vector $\boldsymbol{\eta}$ may be available, in addition to an upper bound on the cardinality of $\boldsymbol{\eta}$. This important problem is referred to as *robust regression* and appears in real-world situations when some observations, named outliers, are completely wrong in an unpredictable way. This could occur during an image acquisition with several corrupted pixels, or result from communication issues during data transmission in sensor networks. Such problems arise in different domains of applications and have been studied in the literature. In the context of electric power grid, the regression problem is known as state estimation, where the goal is to find the operating point of the system based on the voltage signals measured at buses and power signals measured over lines and at buses (Abur and Exposito, 2004; Madani et al., 2017b; Zhang et al., 2018b). Outliers in this case are associated with faulty sensors, cyber attacks, or regional data manipulation to impact the electricity market (Jin et al., 2017; Madani et al., 2017b).

There are several classical works on robust regression and outliers detection. The book by Rousseeuw and Leroy (2005) offers an overview of many fundamental results in this area dating back to 1887 when Edgeworth proposed the least-absolute-value regression estimator. Modern techniques for handling sparse errors of arbitrary magnitudes vary with respect to different features: statistical properties of the error, class of the regression model $f(\mathbf{x};\mathbf{a})$, set of possible true models, type of theoretical guarantees, and characteristics of the adversary model generating errors (Candès et al., 2011; Nasrabadi et al., 2011; Bhatia et al., 2015; Zhang et al., 2016; Klopp et al., 2017). There is a plethora of papers on this topic for the well-known linear regression problem (Candes and Tao, 2005; Wright and Ma, 2010; Studer et al., 2012; Chen et al., 2013a; Bhatia et al., 2017). In this case, the function $f(\mathbf{x};\mathbf{a})$ is linear in the model vector $\mathbf{x}$, and can be written as $\mathbf{a}^*\mathbf{x}$. Nevertheless, there are far less results known for nonlinear regression. This is due to the fact that linear regression amounts to a system of linear equations with a cubic solution complexity if the measurements are error-free, whereas nonlinear regression is NP-hard and its complexity further increases with the inclusion of premeditated errors. However, very special cases of nonlinear regression have been extensively studied in the literature. In particular, the robust phase retrieval problem that can be formulated with $f(\mathbf{x};\mathbf{a}_i) = |\mathbf{a}_i^*\mathbf{x}|^2$ has received considerable attention (Zhang et al., 2016; Hand and Voroninski, 2016; Chen et al., 2017). Another special case is the trace regression problem that has been studied in Hamidi and Bayati (2019) under a low-rank assumption on the unknown matrix solution. However, this has not yet been studied under adversarial sparse additive errors. The mathematical framework provided in the current paper addresses the trace regression problem under a low-rank assumption and sparse adversarial noise.

Given $\mathbf{a} \in \mathcal{A}$ and an arbitrary $\varepsilon > 0$, it follows from the Stone-Weierstrass theorem that there exists a polynomial $p_a : \mathbb{R}^n \to \mathbb{R}$ that uniformly approximates f on $\mathcal{X}$ with the precision error of ε. This way, given the data $\{(y_i, \mathbf{a}_i)\}_{i=1}^m$, there exists a nonlinear regression model

$$y_i = p_{\mathbf{a}_i}(\mathbf{x}) + \hat{\varepsilon}_i, \qquad \forall i \in \{1, ..., m\}$$

where each function $p_{\mathbf{a}_i}(\mathbf{x})$ is a polynomial and $\hat{\varepsilon}_i$ is the difference between $p_{\mathbf{a}_i}(\mathbf{x})$ and $f(\mathbf{x}, \mathbf{a}_i)$ that is bounded from above by ε. Notice that $\hat{\varepsilon}$ is dense noise of a small value

that we do not consider in this paper since its presence just shifts the solutions recovered using our methods by a small value that can be naturally bounded (this corresponds to the sensitivity analysis of conic optimization). On the other hand, each polynomial equation can be converted to a quadratic equation by introducing new variables and adding new quadratic equations (Sojoudi et al., 2014). As an example, the polynomial equation $1 = x^4 - x^3 + x$ can be written as $1 = z^2 - xz + x$ with the additional variable z and measurement equation $0 = z - x^2$ (note that the number of variables and constraints increases in a logarithmic fashion in terms of the degree of the polynomial). This discussion implies that every nonlinear regression could be approximated up to any arbitrary precision with a quadratic regression where the augmented model of the system is quadratic. For this reason, the focus of this paper is only on quadratic regression.

As a far more general case of phase retrieval, a quadratic regression problem with the variable $\mathbf{x}$ can be modeled as $f(\mathbf{x}; \mathbf{A}_i) = \mathbf{x}^* \mathbf{A}_i \mathbf{x}$. The state estimation problem for power systems belongs to the above model due to the quadratic laws of physics (i.e., the quadratic relationship between voltage and power), where each matrix $\mathbf{A}_i$ has rank 1 or 2. Robust regression in power systems is referred to as *bad data detection*. This problem was first studied in 1971 by Merrill and Schweppe (1971), and there are many recent progresses on this topic (Deka et al., 2015; Weng et al., 2015; Madani et al., 2017b).

The existing approaches for robust regression include the analysis of the unconstrained case (Candes and Tao, 2005; Studer et al., 2012; Bhatia et al., 2015, 2017; Josz et al., 2018), the constrained scenario with conditions on the sparsity of the solution vector $\mathbf{x}$ (Wright and Ma, 2010; Nasrabadi et al., 2011; Nguyen and Tran, 2013; McWilliams et al., 2014), and more sophisticated scenarios in the context of matrix completion (Candès et al., 2011; Chen et al., 2013b; Klopp et al., 2017; Zhang et al., 2018a). Motivated by applications in inverse covariance estimation (Wang and Lin, 2014), the papers by Xu et al. (2009); Yang and Xu (2013); McWilliams et al. (2014) consider sparse noise in the input vector $\mathbf{a}_i$ as opposed to the additive error considered in the present paper. The work of Candes and Tao (2005) is based on l_1-minimization, whereas Nasrabadi et al. (2011) solve an extended Lasso formulation defined as the minimization of $\|\mathbf{y} - \mathbf{A}\mathbf{x} + \boldsymbol{\nu}\|_2^2 + \mu_1 \|\mathbf{x}\|_1 + \mu_2 \|\boldsymbol{\nu}\|_1$. The work by Dalalyan and Chen (2012) proposes to solve a second-order cone programming (SOCP) for robust linear regression, which is related to the current paper with a focus on robust nonlinear regression. In contrast to the above-mentioned papers that aim to develop a single optimization problem to estimate the solution of a regression, there are iterative-based methods as well. For instance, Chen et al. (2013a); Bhatia et al. (2015, 2017) propose iterative algorithms via hard thresholding. As a major generalization, the current paper significantly advances the ideas proposed in its conference version (Molybog et al., 2018). Here, we develop an improved theoretical analysis of the semidefinite programming relaxation and provide a more computationally tractable relaxation based on second-order cone programming. This problem can be solved significantly faster than a semidefinite programming of comparable size, which provides the practitioners with a trade-off between the tightness of the relaxation and the computational speed. We put the explanation and the proof into the same framework as the semidefinite relaxation, so that the material from Molybog et al. (2018) can set the stage for presenting the new mathematical results. Aside from theoretical developments, we present a novel set of numerical experiments that

partially answer questions raised by Molybog et al. (2018) and discuss intriguing areas of applications where the more scalable version of the relaxation can make a difference.

Due to the diversity in the problem formulation and approaches taken by different papers, it is difficult to compare the existing results since there is no single dominant method. However, the most common measures of performance for robust regression algorithms are the traditional algorithmic complexity and the permissible number of gross measurements $\|\boldsymbol{\eta}\|_0$ compared to the total number of measurements m. In this paper, the objective is to design a polynomial-time algorithm, in contrast with potentially exponential-time approaches (Víšek, 2006), with guaranteed convergence under technical assumptions. As far as the robustness of an algorithm is concerned, the existing works often provide probabilistic guarantees on the recoverability of the original parameter vector $\mathbf{x}$ for linear Gaussian stochastic systems under various assumptions on the relationship between $\|\boldsymbol{\eta}\|_0$ and m. In this case, the ratio $\frac{\|\boldsymbol{\eta}\|_0}{m}$, named breakdown point, is limited by a constant and could even approach 1 if the unknown solution $\mathbf{x}$ is sparse.

1.1 Contributions and Organization

The main objective of this paper is to analyze a robust regression problem for an arbitrary quadratic model that includes power system state estimation and phase retrieval as special cases. The focus is on the calculation of the maximum number of bad measurements that does not compromise the exact reconstruction of the model vector $\mathbf{x}$. In Section 2, we formally state the problem. In Section 3, we propose two conic optimization methods and study their properties. In particular, we obtain conditions that guarantee the exact reconstruction of $\mathbf{x}$. In Section 4, we develop the main results of this paper. Under certain technical assumptions, we discover the dependence between the number of perfect measurements and the maximum admissible number of wrong measurements. After that, we consider a stochastic setting based on Gaussian distributions. In this case, we show that the number of bad measurements can safely be on the order of the square root of the total number of measurements, and moreover the breakpoint approaches 1/2 if there is enough prior information. To provide a broader range of possible approaches to the problem, Section 5 designs an alternative iterative-based method. Numerical results are presented in Section 6, which includes a case study on a European power grid.

1.2 Notation

$\mathbb{R}^n$ and $\mathbb{C}^n$ denote the sets of real and complex n-dimensional vectors, respectively. Bold letters are reserved for vectors and matrices. $[\mathbf{A}]_{ij}$ or A_{ij} is the (i,j)-th element of a matrix $\mathbf{A}$. The symbols $\mathbb{H}^n$ and $\mathbb{S}^n$ denote the sets of $n \times n$ Hermitian and symmetric matrices. $\mathrm{tr}(\mathbf{A})$ and $\langle \mathbf{A}, \mathbf{B} \rangle$ are the trace of a matrix $\mathbf{A}$ and the Frobenius inner product of two matrices $\mathbf{A}$ and $\mathbf{B}$. The conjugate transpose and Moore-Penrose pseudoinverse of $\mathbf{A}$ are shown as $\mathbf{A}^*$ and $\mathbf{A}^+$. The notation $\mathbf{A} \circ \mathbf{B}$ refers to the Hadamard (entrywise) multiplication. The eigenvalues of a matrix $\mathbf{M} \in \mathbb{H}^n$ are denoted as $\lambda_1(\mathbf{M}), ..., \lambda_n(\mathbf{M})$ in descending order. The smallest and the largest singular values of $\mathbf{A}$ are shown as $\sigma_{\min}$ and $\sigma_{\max}$, respectively. $\mathbf{e}_i$ stands for the i-th column of the unit matrix $\mathbf{I}$ of appropriate dimension. Given a matrix $\mathbf{A} \in \mathbb{C}^{n \times m}$ and a set $\mathcal{S} \subset \{1, \ldots, m\}$, the matrix $\mathbf{A}_{\mathcal{S}}$ is defined to be a matrix obtained by adjoining the columns of $\mathbf{A}$ with indexes in $\mathcal{S}$. Given a vector $\mathbf{a} \in \mathbb{C}^n$ and a

set $\mathcal{S} \subset \{1, \ldots, n\}$, the vector $\mathbf{a}_{\mathcal{S}}$ is defined to be a subvector of $\mathbf{a}$ obtained by stacking the components of $\mathbf{a}$ with indexes in $\mathcal{S}$. For a sequence of indexes $\mathcal{S}$, the symbol $\{\alpha_i\}_{i\in\mathcal{S}}$ denotes a sequence indexed by $\mathcal{S}$. Whenever the notation is obvious from the context, we drop the indexing subscript for notational simplicity. The symbol $\|\mathbf{v}\|_0$ shows the cardinality of a vector $\mathbf{v}$, i.e., the number of its nonzero elements. Given a matrix $\mathbf{A}$, the symbols $\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_\infty$, $\|\mathbf{A}\|_2$, and $\|\mathbf{A}\|_F$ denote the maximum absolute column sum, maximum absolute row sum, maximum singular value, and Frobenius norm of $\mathbf{A}$, respectively. The cardinality of a set $\mathcal{M}$ is indicated as $|\mathcal{M}|$. The notation $a \sim \mathcal{N}(\alpha, \beta)$ means that a is a normally distributed random variable with the parameters α and β.

2. Problem Formulation and Preliminaries

The quadratic regression under sparse noise aims to find a vector $\mathbf{x}$ in $\mathbb{R}^n$ or $\mathbb{C}^n$ such that

$$y_r = \mathbf{x}^*\mathbf{M}_r\mathbf{x} + \eta_r, \quad \forall r \in \{1, \ldots, m\}, \tag{2}$$

where

- $y_1, \ldots, y_m$ are some known real-valued measurements.
- $\eta_1, \ldots, \eta_m$ are unknown but sparsely occurring real-valued noise of arbitrary magnitudes.
- $\mathbf{M}_1, \ldots \mathbf{M}_m$ are some known $n \times n$ Hermitian matrices.

The regression problem could have two solutions $\pm\mathbf{x}$ in the real-valued case, which increases to infinitely many in the form of $\mathbf{x} \times e^{\sqrt{-1}\theta}$ in the complex case. To avoid this ambiguity, the objective of this work is to find the matrix $\mathbf{x}\mathbf{x}^*$ rather than $\mathbf{x}$ since this matrix is invariant under the rotation of $\mathbf{x}$. At the same time, the recovery of $\mathbf{x}$ from $\mathbf{x}\mathbf{x}^*$ is a simple problem that can be solved using the spectral decomposition. If m is large enough, then $\mathbf{x}\mathbf{x}^*$ is expected to be unique. This paper aims to recover any solution $\mathbf{x}\mathbf{x}^*$ in case there are multiple ones. In Problem (2), each measurement equation could have a linear term in addition to its purely quadratic function $\mathbf{x}^*\mathbf{M}_r\mathbf{x}$. By introducing one additional variable z such that $z^2 = 1$, one can multiply the linear terms with z to make them quadratic (Madani et al., 2017c). As a result, Problem (2) is a general quadratic regression problem.

Let $\hat{\mathbf{x}}$ be an initial guess for the unknown solution $\mathbf{x}$. We refer to this as *prior knowledge*. We do not make any assumption about the gap between $\hat{\mathbf{x}}$ and $\mathbf{x}$, and develop different methods that can be run independent of this gap. However, the goal is to show that as this gap becomes smaller, the performance of these methods increases. More precisely, we define a measure to quantify the amount of information in the prior knowledge and use it to study the to-be-developed techniques. Note that it is easy to deduce prior knowledge for many real-world systems. For example, we will show that the physics of power systems naturally provide such useful knowledge.

Consider the complex-valued case and write $\mathbf{x} = \mathbf{a} + \sqrt{-1}\mathbf{b} \in \mathbb{C}^n$ and $\mathbf{M} = \mathbf{A} + \sqrt{-1}\mathbf{B} \in \mathbb{H}^n$, where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{S}^n$ and $\mathbf{B} = -\mathbf{B}^T \in \mathbb{R}^{n\times n}$. It is straightforward to verify that:

$$\mathbf{x}^*\mathbf{M}\mathbf{x} = \mathbf{a}^T\mathbf{A}\mathbf{a} + 2\mathbf{b}^T\mathbf{B}\mathbf{a} + \mathbf{b}^T\mathbf{A}\mathbf{b} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}^T \begin{bmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}$$

Notice that the matrices in the right-hand side of the above equation are real-valued. As a result, we will only develop the theoretical results of this work in the real-valued case $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{M}_r \in \mathbb{S}^n$ since they can be easily carried over to the complex-valved case. However, we will offer a case study on power systems where the unknown state is a complex vector.

In the regression problem under sparse noise, the vector $\boldsymbol{\eta}$ is assumed to be sparse. To distinguish between error-free and erroneous measurements, we partition the set of measurements into two subsets of *good* and *bad* measurements:

$$\mathcal{G} = \{r \in \{1, \dots, m\} | \eta_r = 0\}, \quad \mathcal{B} = \{1, \dots, m\} \backslash \mathcal{G}$$

To streamline the derivation of the analytical results of this paper, we assume that $\mathcal{G} = \{1, \dots, |\mathcal{G}|\}$ and $\mathcal{B} = \{|\mathcal{G}|+1, \dots, m\}$. However, the algorithms to be designed are completely oblivious to the type of each measurement and its membership in either $\mathcal{G}$ or $\mathcal{B}$.

The objective of this paper is to develop efficient algorithms for finding $\mathbf{x}$ precisely as long as $\boldsymbol{\eta}$ is sufficiently sparse. This statement will be formalized in the next sections.

3. Conic Optimization Methods

Consider a variable matrix $\mathbf{W}$ playing the role of $\mathbf{x}\mathbf{x}^T$. This matrix is positive semidefinite and has rank 1. By dropping the rank constraint, one can cast the quadratic regression as a linear matrix regression. Motivated by this relaxation, consider the optimization problem

$$\begin{aligned} \underset{\mathbf{W} \in \mathbb{S}^n,\ \boldsymbol{\nu} \in \mathbb{R}^m}{\text{minimize}} \quad & \langle \mathbf{W}, \mathbf{M} \rangle + \mu \|\boldsymbol{\nu}\|_1 \\ \text{subject to} \quad & \langle \mathbf{W}, \mathbf{M}_r \rangle + \nu_r = y_r, \quad \forall r \in \{1, \dots, m\} && \text{(3a)} \\ & \mathbf{W} = \mathbf{W}^T \succeq_{\mathcal{C}} 0 && \text{(3b)} \end{aligned}$$

where the notation $\succeq_{\mathcal{C}}$ is the generalized inequality sign with respect to $\mathcal{C}$, which is either the cone of symmetric positive semidefinite (PSD) matrices or the 2×2 principal sub-matrices PSD cone (see Permenter and Parrilo, 2014). The above cones are formally defined in Subsection 3.2.

The problem definition involves a matrix $\mathbf{M}$ that is to be designed based on the prior knowledge $\hat{\mathbf{x}} \in \mathbb{R}^n$ in such a way that the term $\langle \mathbf{W}, \mathbf{M} \rangle$ in the objective function promotes a low-rank structure on $\mathbf{W}$. The construction of $\mathbf{M}$ will be studied later in the paper. We refer to Problem (3) as *penalized conic program*, but call it with more specific names in two special cases: (i) *penalized semidefinite program (SDP)* if $\mathcal{C}$ is the cone of PSD matrices, (ii) *penalized second-order cone program (SOCP)* if $\mathcal{C}$ is the cone of matrices with all 2×2 principal sub-matrices being PSD. The penalized conic program is a convex problem and can be solved in polynomial time up to any given accuracy.

A popular approach to solving rank minimization problems is via an approximation technique that replaces the non-convex objective function with the nuclear norm of the unknown matrix (Candès et al., 2011). We exploit a different approach for three main reasons:

- The nuclear norm minimization is rooted in the fact that the nuclear norm is a convex envelope of the rank over a certain ball, but the connection between nuclear norm and rank fades away when the ball is intersected with the hyperplanes given by (3a).

- In many practical applications, some prior knowledge about the unknown state is available. However, the nuclear norm minimization cannot incorporate such information to improve the search for the unknown solution. This is in contrary to the standard numerical algorithms for optimization that allow the initialization of the process for finding an optimal solution. Therefore, one would expect to have a new learning method for quadratic regression that exploits prior knowledge about the solution.

- The minimization of the trace is meaningless in many applications where the trace of all feasible matrices $\mathbf{W}$ is automatically in a narrow bound. In this case, the trace cannot be used to distinguish low-rank solutions from high-rank solutions. This naturally occurs in power systems, for which the trace is almost fixed since voltage magnitudes are always close to nominal values (e.g., 110 volts) (Madani et al., 2014)

The method to be developed in this paper addresses the above issues via a major generalization of the nuclear norm minimization. In particular, if $\hat{\mathbf{x}} = 0$, then the proposed approach is equivalent to the nuclear norm minimization. We refer to $\hat{\mathbf{x}}$ as prior knowledge, and aim to show how the amount of information in the prior knowledge—measured in terms of the closeness between $\hat{\mathbf{x}}$ and the unknown solution—affects the performance of the penalized conic program and the estimation error.

In the following two subsections, we will introduce the functions κ and ξ, matrices $\bar{\mathbf{J}}$ and $\tilde{\mathbf{J}}$, and vectors $\bar{\mathbf{d}}$ and $\tilde{\mathbf{d}}$, and then study the problem of designing $\mathbf{M}$ based on the prior knowledge $\hat{\mathbf{x}}$.

3.1 Penalized Semidefinite Programming

Consider the penalized SDP that corresponds to Problem (3) with $\mathcal{C}$ equal to the PSD cone. Given a matrix $\mathbf{X} \in \mathbb{S}^n$, define $\kappa(\mathbf{X})$ to be the sum of the two smallest eigenvalues of $\mathbf{X}$, i.e.,

$$\kappa(\mathbf{X}) := \lambda_n(\mathbf{X}) + \lambda_{n-1}(\mathbf{X})$$

Let the matrix $\mathbf{M}$ in the objective function of the penalized SDP be chosen to have the following properties:

$$\begin{aligned} &\mathbf{M}\hat{\mathbf{x}} = \mathbf{0}, \\ &\operatorname{rank}(\mathbf{M}) \geq n-1 \\ &\kappa(\mathbf{M}) > 0 \end{aligned}$$

If there is no prior knowledge available, one can select $\hat{\mathbf{x}}$ to be zero and then choose $\mathbf{M}$ as $\mathbf{I}$. This will correspond to the famous nuclear norm minimization. As will become evident in the paper, the linear term $\langle \mathbf{W}, \mathbf{M} \rangle$ with the above-mentioned matrix $\mathbf{M}$ penalizes the deviation of $\mathbf{W}$ from $\hat{\mathbf{x}}\hat{\mathbf{x}}^T$. Since $\hat{\mathbf{x}}\hat{\mathbf{x}}^T$ is low-rank, the inclusion of this linear term automatically takes care of both prior knowledge and low-rank promotion. There are infinitely many choices for $\mathbf{M}$, and it is not important which one to select as far as the analysis of this paper is concerned. One natural choice for $\mathbf{M}$ is the matrix of orthogonal projection onto the hyperplane that is orthogonal to $\hat{\mathbf{x}}$. This particular matrix is computationally cheap to construct. However, if more than one initial guess is available, it is beneficial to design the matrix $\mathbf{M}$ via an optimization problem that attempts to minimize the violation of the above conditions for all initial values of $\hat{\mathbf{x}}$.

Observe that the dual of Problem (3) can be obtained as:

$$\begin{aligned} \underset{\lambda \in \mathbb{R}^m}{\text{maximize}} \quad & -\mathbf{y}^T \boldsymbol{\lambda} \\ \text{subject to} \quad & \mathbf{M} + \sum_{r=1}^{m} \lambda_r \mathbf{M}_r \succeq 0 \qquad (4a) \\ & \|\boldsymbol{\lambda}\|_\infty \leq \mu \qquad (4b) \end{aligned}$$

where $\succeq 0$ is the positive semidefinite sign. Define the matrix $\bar{\mathbf{J}}$ and the vector $\bar{\mathbf{d}}$ as:

$$\bar{\mathbf{J}} = [\mathbf{M}_1 \mathbf{x} \ \ldots \ \mathbf{M}_m \mathbf{x}] \qquad (5a)$$

$$\bar{\mathbf{d}} = \mathbf{M}\mathbf{x} \qquad (5b)$$

where $\mathbf{x}$ is the solution of the original problem (2). The matrix $\bar{\mathbf{J}}$ captures the coherence between the model vector $\mathbf{x}$ and the measurement matrices $\mathbf{M}_r$. At its turn, the vector $\bar{\mathbf{d}}$ measures the alignment of the solution $\mathbf{x}$ and the prior knowledge $\hat{\mathbf{x}}$. Note that $\bar{\mathbf{J}}$ and $\bar{\mathbf{d}}$ are both completely noise-agnostic. The regularity property of the matrix $\bar{\mathbf{J}}$ and the norm of the vector $\bar{\mathbf{d}}$ play important roles in guaranteeing the correct recovery of $\mathbf{x}$. A preliminary result is provided below, which will later be used to study the penalized SDP.

Lemma 1 *Assume that there exists an index $r \in \{1, \ldots, m\}$ such that $\hat{\mathbf{x}}^T \mathbf{M}_r \hat{\mathbf{x}} \neq 0$ and*

$$\mu > \|\bar{\mathbf{J}}_{\mathcal{G}}^{+} \left(\bar{\mathbf{d}} - \mu \bar{\mathbf{J}}_{\mathcal{B}} \ \text{sign}(\boldsymbol{\eta}_{\mathcal{B}}) \right) \|_\infty \qquad (6a)$$

$$\frac{\kappa(\mathbf{M})}{2 \max_r \|\mathbf{M}_r\|_2} > \|\bar{\mathbf{J}}_{\mathcal{G}}^{+} \left(\bar{\mathbf{d}} - \mu \bar{\mathbf{J}}_{\mathcal{B}} \ \text{sign}(\boldsymbol{\eta}_{\mathcal{B}}) \right) \|_1 + \mu |\mathcal{B}| \qquad (6b)$$

Then, $(\mathbf{x}\mathbf{x}^T, \boldsymbol{\eta})$ is the unique solution of the penalized SDP. Moreover, $\hat{\boldsymbol{\lambda}} = \begin{bmatrix} \hat{\boldsymbol{\lambda}}_{\mathcal{G}} \\ \hat{\boldsymbol{\lambda}}_{\mathcal{B}} \end{bmatrix}$ defined as

$$\begin{aligned} \hat{\boldsymbol{\lambda}}_{\mathcal{B}} &= -\mu \ \text{sign}(\boldsymbol{\eta}_{\mathcal{B}}) \\ \hat{\boldsymbol{\lambda}}_{\mathcal{G}} &= -\bar{\mathbf{J}}_{\mathcal{G}}^{+} \left(\bar{\mathbf{d}} + \bar{\mathbf{J}}_{\mathcal{B}} \hat{\boldsymbol{\lambda}}_{\mathcal{B}} \right) \end{aligned}$$

is a dual solution.

Proof The proof is provided in Appendix A. ■

The conditions given in Lemma 1 will be refined and further studied in Section 4 to uncover useful properties of the penalized SDP.

3.2 Penalized Second-Order Cone Programming

Although penalized SDP is a convex optimization, its memory and time complexities make it less appealing for large-scale problems (Boyd et al., 2004). These complexities can be

significantly reduced if the union of the 0-1 sparsity patterns of the matrices $\mathbf{M}, \mathbf{M}_1, ..., \mathbf{M}_m$ is a sparse matrix itself (Fukuda et al., 2001). This requires a natural sparsity in the measurement matrices $\mathbf{M}_r$ and also the design of a sparse matrix $\mathbf{M}$, which is not always possible. As an alternative, one can break down the complexity of the penalized SDP by replacing its constraint $\mathbf{W} \succeq 0$ with second-order conic constraints. Although penalized SDP offers better recovery guarantees than penalized SOCP, the latter has a significantly lower computational complexity and can be efficiently solved for large-scale problems using interior-point methods (Alizadeh and Goldfarb, 2003). In this part, we study the penalized SOCP as a counterpart of penalized SDP. This optimization problem is obtained by building the cone $\mathcal{C}$ based on the 2×2 principal submatrices of $\mathbf{W}$, as explained below.

Definition 2 ($2\mathcal{PSM}$) *A matrix $\mathbf{X} \in \mathbb{S}^n$ belongs to the 2×2 principal sub-matrices PSD cone if each 2×2 principal sub-matrix of $\mathbf{X}$ is positive semidefinite, i.e.,*

$$[\mathbf{e}_i \;\; \mathbf{e}_j]^T \mathbf{X} [\mathbf{e}_i \;\; \mathbf{e}_j] \succeq 0, \qquad \forall\, i < j$$

Since the $2\mathcal{PSM}$ cone is not self-dual, we introduce the scaled diagonally dominant cone below.

Definition 3 ($\mathcal{SDD}$) *A matrix $\mathbf{X} \in \mathbb{R}^{n \times n}$ belongs to the scaled diagonally dominant cone if there exists a set of 2×2 positive semidefinite matrices $\{\mathbf{X}^{ij}\}_{i<j}^{j \leq n}$ such that*

$$\sum_{j=2}^{n} \sum_{i=1}^{j-1} [\mathbf{e}_i \;\; \mathbf{e}_j] \mathbf{X}^{ij} [\mathbf{e}_i \;\; \mathbf{e}_j]^T = \mathbf{X}$$

The notation $\{\mathbf{X}^{ij}\}_{i<j}^{j \leq n}$ in the above definition means $\{\mathbf{X}^{ij} \big| j = 2, \ldots, n, i = 1, \ldots, j-1\}$. The next lemma explains the connection between $2\mathcal{PSM}$ and $\mathcal{SDD}$ cones.

Lemma 4 (Permenter and Parrilo (2014)) *The dual of the 2×2 principal sub-matrices PSD cone is the scaled diagonally dominant cone of the same dimension.*

In what follows, we will define and describe certain properties of a linear space of diagonal decompositions of matrices. These definitions are somewhat more tedious than the ones in the previous subsection, but they serve the same aim: they formally define the matrix $\mathbf{M}$ and the counterparts of $\bar{\mathbf{J}}$ and $\bar{\mathbf{d}}$ for the penalized SOCP.

Definition 5 *The sequence $\{\mathbf{A}^{ij} \in \mathbb{S}^2\}_{i<j}$ is said to be a diagonal decomposition (or just decomposition) of $\mathbf{A} \in \mathbb{S}^n$ if*

$$\mathbf{A} = \sum_{j=2}^{n} \sum_{i=1}^{j-1} [\mathbf{e}_i, \mathbf{e}_j] \mathbf{A}^{ij} [\mathbf{e}_i, \mathbf{e}_j]^T$$

A decomposition that consists of PSD matrices is a certificate that a matrix belongs to the $\mathcal{SSD}$ cone. Similarly to the function κ defined for the penalized SDP, we introduce the function $\chi(\{\mathbf{X}^{ij}\}_{i<j})$ as follows:

$$\chi\left(\{\mathbf{X}^{ij}\}_{i<j}\right) := \min_{i<j} \operatorname{tr}(\mathbf{X}^{ij}) = \min_{i<j} \left(\lambda_1(\mathbf{X}^{ij}) + \lambda_2(\mathbf{X}^{ij})\right)$$

Consider a sequence $\{\mathbf{M}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ such that

$$\begin{cases} \chi(\{\mathbf{M}^{ij}\}_{i<j}) > 0 \\ \mathbf{M}^{ij}[\hat{x}_i \;\; \hat{x}_j]^T = \mathbf{0} \;\; \text{for all } i < j \end{cases}$$

Define the corresponding penalized SOCP as Problem (3) with $\mathcal{C}$ equal to the $2\mathcal{PSM}$ cone and

$$\mathbf{M} = \sum_{i<j} [\mathbf{e}_i \;\; \mathbf{e}_j]\mathbf{M}^{ij}[\mathbf{e}_i \;\; \mathbf{e}_j]^T. \tag{7}$$

Since $\mathbf{M}^{ij} \succeq 0$, the matrix $\mathbf{M}$ belongs to the $\mathcal{SDD}$ cone. Similarly to the penalized SDP, there is an infinite number of possible matrices $\mathbf{M}$, one of which can naturally be obtained by selecting $\mathbf{M}^{ij}$ to be the orthogonal projection onto the line orthogonal to $[\hat{x}_i \;\; \hat{x}_j]^T$.

The dual of the penalized SOCP takes the form:

$$\begin{aligned} \text{maximize} \quad & -\mathbf{y}^T\boldsymbol{\lambda} \\ \text{subject to} \quad & \mathbf{M} + \sum_{r=1}^{m} \lambda_r \mathbf{M}_r = \mathbf{H} && \text{(8a)} \\ & \sum_{i<j} [\mathbf{e}_i \;\; \mathbf{e}_j]\mathbf{H}^{ij}[\mathbf{e}_i \;\; \mathbf{e}_j]^T = \mathbf{H} && \text{(8b)} \\ & \mathbf{H}^{ij} \succeq 0 && \text{(8c)} \\ & \|\boldsymbol{\lambda}\|_\infty \leq \mu && \text{(8d)} \end{aligned}$$

where the variables are $\lambda \in \mathbb{R}^m$, $\mathbf{H} \in \mathbb{S}^n$ and $\{\mathbf{H}^{ij}\}_{i<j}^{j\leq n} \subset \mathbb{S}^2$. Now, it is easy to observe that each conic constraint $[\mathbf{e}_i \;\; \mathbf{e}_j]^T\mathbf{W}[\mathbf{e}_i \;\; \mathbf{e}_j] \succeq 0$ in Problem (3) corresponds to the dual variable matrix $\mathbf{H}^{ij}$. Hence, the complementary slackness condition can be written as

$$\langle [\mathbf{e}_i \;\; \mathbf{e}_j]^T\mathbf{W}[\mathbf{e}_i \;\; \mathbf{e}_j], \mathbf{H}^{ij} \rangle = 0, \qquad \text{for all } i < j \leq n$$

Define $\mathbf{G}$ to be a symmetric matrix such that $\mathbf{M}^{ij}[x_i \;\; x_j]^T = [G_{ij} \;\; G_{ji}]^T$ for all $i < j \in \{1, \ldots, n\}$ and $G_{ii} = 0$ for all $i \in \{1, \ldots, n\}$. Furthermore, for every $r \in \{1, \ldots m\}$, define $\mathbf{R}^r$ as a matrix with the properties:

$$\begin{cases} \sum\limits_{j=1}^{n} R_{ij}^r = M_{ii}^r \;\; \text{for all } i \in \{1, \ldots, n\} \\ R_{ii}^r = 0 \;\; \text{for all } i \in \{1, \ldots, n\} \end{cases}$$

One simple example of this matrix is a matrix with the (i,j)-entry equal to $R_{ij}^r = \frac{M_{ii}^r}{n-1}$ for $i \neq j$. Given $r \in \{1, \ldots, m\}$, define $\mathbf{G}^r \in \mathbb{R}^{n\times n}$ as a matrix with the components $G_{ij}^r = x_j M_{ij}^r + x_i R_{ij}^r$ and $G_{ii}^r = 0$ for all $i, j \in \{1, \ldots, n\}$. Similarly to (5), define:

$$\begin{aligned} \tilde{\mathbf{J}} = & \;\; \left[\;\text{vecnd}\left(\mathbf{G}^1\right) \;\; \ldots \;\; \text{vecnd}\left(\mathbf{G}^m\right) \;\right] && \text{(9a)} \\ \tilde{\mathbf{d}} = & \;\; \text{vecnd}\left(\mathbf{G}\right) && \text{(9b)} \end{aligned}$$

where the vectorization operator $\text{vecnd} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n^2-n}$ puts all elements excluding the diagonal of its matrix argument into the form of a vector. Similarly to the penalized SDP

case, here $\tilde{\mathbf{J}}$ captures the coherence between the data and the true model, while $\tilde{\mathbf{d}}$ captures the correlation between the true model and the prior knowledge.

The counterpart of Lemma 1 is stated below for the penalized SOCP.

Lemma 6 *Assume that the components of the initial guess are nonzero (i.e., $\hat{x}_i \neq 0$ for all $i \in \{1,\dots,n\}$) and that there exists an index $r \in \{1,\dots,m\}$ such that $\hat{\mathbf{x}}^*\mathbf{M}_r\hat{\mathbf{x}} \neq 0$ and*

$$\mu > \|\tilde{\mathbf{J}}_{\mathcal{G}}^{+}\left(\tilde{\mathbf{d}} - \mu\tilde{\mathbf{J}}_{\mathcal{B}}\ \mathrm{sign}(\boldsymbol{\eta}_{\mathcal{B}})\right)\|_\infty \tag{10a}$$

$$\frac{\chi(\{\mathbf{M}^{ij}\}_{i<j})}{\max\limits_{r,\ i<j}|\mathrm{tr}(\{\mathbf{M}_r^{ij}\}_{i<j})|} > \|\tilde{\mathbf{J}}_{\mathcal{G}}^{+}\left(\tilde{\mathbf{d}} - \mu\tilde{\mathbf{J}}_{\mathcal{B}}\ \mathrm{sign}(\boldsymbol{\eta}_{\mathcal{B}})\right)\|_1 + \mu|\mathcal{B}| \tag{10b}$$

Then, $(\mathbf{x}\mathbf{x}^T, \boldsymbol{\eta})$ is the unique solution of the penalized SOCP. Moreover, $\hat{\boldsymbol{\lambda}} = \begin{bmatrix} \hat{\boldsymbol{\lambda}}_{\mathcal{G}} \\ \hat{\boldsymbol{\lambda}}_{\mathcal{B}} \end{bmatrix}$ defined as

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{\mathcal{B}} &= -\,\mu\ \mathrm{sign}(\boldsymbol{\eta}_{\mathcal{B}})\\ \hat{\boldsymbol{\lambda}}_{\mathcal{G}} &= -\,\tilde{\mathbf{J}}_{\mathcal{G}}^{+}\left(\tilde{\mathbf{d}} + \tilde{\mathbf{J}}_{\mathcal{B}}\hat{\boldsymbol{\lambda}}_{\mathcal{B}}\right)\end{aligned}$$

can be completed to a dual optimal solution.

Proof The proof is provided in Appendix B. ■

We need to mention that while there is some freedom in the choice of $\hat{\boldsymbol{\lambda}}_{\mathcal{G}}$ in the proof of Lemma 6, the dual variables $\hat{\boldsymbol{\lambda}}_{\mathcal{B}}$ associated with the bad measurements are inflexible. This is elaborated below.

Lemma 7 *$\hat{\boldsymbol{\lambda}}_{\mathcal{B}} = -\mu\ \mathrm{sign}(\boldsymbol{\eta}_{\mathcal{B}})$ is the only possible choice for the optimal dual variables if the optimal primal variables are $(\mathbf{x}\mathbf{x}^T, \boldsymbol{\eta})$.*

Proof The proof is provided in Appendix B ■

4. Main Results

In this section, we develop the key theoretical results on the Robust Quadratic Regression solution via the conic methods presented in the preceding section. The common structure of the conditions in Lemmas 1 and 6 allows us to derive results providing guarantees for both the SDP and the SOCP approaches simultaneously. To do so, we will use the universal notations $\mathbf{J}$ and $\mathbf{d}$ to denote $\bar{\mathbf{J}}$ and $\bar{\mathbf{d}}$ (defined in Subsection 3.1) in the penalized SDP case and to denote $\tilde{\mathbf{J}}$ and $\tilde{\mathbf{d}}$ (defined in Subsection 3.2) in the penalized SOCP case. Define

$$\alpha_{\mathrm{SDP}} = \frac{\kappa(\mathbf{M})}{2\|\bar{\mathbf{d}}\|_2 \max\limits_r \|\mathbf{M}_r\|_2} \quad \text{or} \quad \alpha_{\mathrm{SOCP}} = \frac{\chi(\{\mathbf{M}^{ij}\}_{i<j})}{\|\tilde{\mathbf{d}}\|_2 \max\limits_{r,\ i<j}|\mathrm{tr}(\mathbf{M}_r^{ij})|}$$

For the particular matrices $\mathbf{M}$ constructed in Section 3 using the projection operator, both $\kappa(\mathbf{M})$ and $\chi(\{\mathbf{M}^{ij}\}_{i<j})$ are equal to 1. In addition, one can normalize the equations in (2) before solving the problem via a rescaling so that $\|\mathbf{M}_r\| = 1$, in which case the terms with $\mathbf{M}_r$

in the definitions of α_{SDP} and α_{SOCP} can be eliminated (or bounded by a constant). Therefore, we can write that $\alpha_{\text{SDP}} \propto |\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\hat{\mathbf{x}}}{\|\hat{\mathbf{x}}\|} \rangle|^{-1}$ and $\alpha_{\text{SOCP}} \propto (\sum_{i<j} |\langle [\frac{x_i}{\|\mathbf{x}\|} \ \frac{x_j}{\|\mathbf{x}\|}], [\frac{\hat{x}_i}{\|\hat{\mathbf{x}}\|} \ \frac{\hat{x}_j}{\|\hat{\mathbf{x}}\|}] \rangle|)^{-1}$, which imply that these parameters measure the amount of information in the prior knowledge. Henceforth, we use the shorthand notation α to denote α_{SDP} or α_{SOCP} depending on whether the penalized SDP or SOCP is analyzed. The same notation is used for l that takes one of the following values:

$$l_{\text{SDP}} = n; \qquad l_{\text{SOCP}} = n^2 - n$$

4.1 Deterministic Bound

In this subsection, we establish a uniform bound on the number of bad measurements that a penalized conic relaxation can tolerate. To do so, we make use of two matrix properties introduced in Bhatia et al. (2017).

Definition 8 (*SSC property) A matrix $\mathbf{X} \in \mathbb{R}^{l \times m}$ is said to satisfy the Subset Strong Convexity (SSC) Property at level p with constant $\gamma_p > 0$ if*

$$\gamma_p \leq \min_{|S|=p} \sqrt{\lambda_{min}(\mathbf{X}_S \mathbf{X}_S^T)}$$

Definition 9 (*SSS property) A matrix $\mathbf{X} \in \mathbb{R}^{l \times m}$ is said to satisfy the Subset Strong Smoothness (SSS) Property at level p with constant $\Gamma_p > 0$ if*

$$\max_{|S|=p} \sqrt{\lambda_{max}(\mathbf{X}_S \mathbf{X}_S^T)} \leq \Gamma_p$$

Note that the notation $|S| = p$ in the above definition specifies the index set of any p columns of the matrix $\mathbf{X}$. The ratio of the constants γ_p and Γ_{m-p} can be interpreted as a uniform condition number at level p.

Theorem 10 *Consider Problem* (2)*, and let $N = |\mathcal{B}|$ denote the cardinality of the support of the noise vector η. Without any future assumption on the noise, consider the corresponding penalized conic problem. Consider arbitrary constants $\bar{\alpha}$, γ and Γ such that*

$$\bar{\alpha} > \frac{\Big(\sqrt{N}\frac{\Gamma}{\gamma} + (1 - \frac{\Gamma}{\gamma})\Big)\sqrt{m-N} + N}{\gamma - \Gamma},$$

- *If the exact solution $\mathbf{x}$ of* (2)*, the prior knowledge $\hat{\mathbf{x}}$ and the measurement matrices $\mathbf{M}_r$ are such that $\bar{\mathbf{J}}$ satisfies the SSC property at level $m - N = |\mathcal{G}|$ with the constant γ and the SSS property at level $N = |\mathcal{B}|$ with the constant Γ, then there exists a constant μ for which $(\mathbf{x}\mathbf{x}^T, \boldsymbol{\eta})$ is the unique solution of the penalized SDP problem if $\alpha_{SDP} \geq \bar{\alpha}$.*

- *If the exact solution $\mathbf{x}$ of* (2)*, the prior knowledge $\hat{\mathbf{x}}$ and the measurement matrices $\mathbf{M}_r$ are such that $\tilde{\mathbf{J}}$ satisfies the SSC property at level $m - N = |\mathcal{G}|$ with the constant γ and the SSS property at level $N = |\mathcal{B}|$ with the constant Γ, then there exists a constant μ for which $(\mathbf{x}\mathbf{x}^T, \boldsymbol{\eta})$ is the unique solution of the penalized SOCP problem if $\alpha_{SOCP} \geq \bar{\alpha}$.*

Proof The proof follows from Lemmas 1 and 6, together with Lemma 11 to be stated later in the paper. ■

Theorem 10 implies that the penalized conic relaxations are exact and the corresponding instances of the $\mathcal{NP}$-hard problem (2) can be solved in polynomial time, provided that they satisfy a certain condition. This condition is not restrictive as long as the amount of information in the prior knowledge is not too low.

Notice that aside from the cardinalities of the sets $\mathcal{G}$ and $\mathcal{B}$, Theorem 10 imposes no condition on the noise values. Therefore, the guarantee provided by this Theorem is established for the "worst-case scenario" when the adversary is adaptive and strategically selects the indexes of the error vector $\boldsymbol{\eta}$ based on the true solution $\mathbf{x}$ to have the most impact. Theorem 10 in the paper is based on two basic assumptions:

- Incoherence of the good measurements and dominance of good measurements over bad ones: This is implied by the terms of the form $\gamma - \Gamma$ in the inequality bound;
- The amount of information in the prior knowledge: This is implied by the lower bound $\alpha \geq \bar{\alpha}$.

Although Theorem 10 just states the existence of the hyperparameter μ, we will identify an interval for this parameter below.

Lemma 11 *Let $\mathbf{J}$ be a matrix in $\mathbb{R}^{l\times m}$ that satisfies the SSC and SSS properties on levels $|\mathcal{G}|$ and $|\mathcal{B}|$ with the respective constants $\gamma_{|\mathcal{G}|}$ and $\Gamma_{|\mathcal{B}|}$ ($\gamma_{|\mathcal{G}|} > \Gamma_{|\mathcal{B}|}$). Moreover, let $\mathbf{d}$ be a vector in $\mathbb{R}^l$ and $\boldsymbol{\lambda}$ be a vector in $\mathbb{R}^m$ such that $\boldsymbol{\lambda}_{\mathcal{B}} = \mu \cdot \mathbf{s}$, where μ is a scalar and the entries of $\mathbf{s}$ are only $+1$ or -1. If*

$$\alpha\gamma_{|\mathcal{G}|}(\gamma_{|\mathcal{G}|} - \Gamma_{|\mathcal{B}|}) - |\mathcal{B}|\gamma_{|\mathcal{G}|} > \Big(\sqrt{|\mathcal{B}|}\Gamma_{|\mathcal{B}|} + (\gamma_{|\mathcal{G}|} - \Gamma_{|\mathcal{B}|})\Big)\sqrt{|\mathcal{G}|}$$

then the interval

$$\left[\frac{\|\mathbf{d}\|_2}{\gamma_{|\mathcal{G}|} - \Gamma_{|\mathcal{B}|}}, \frac{(\alpha\gamma_{|\mathcal{G}|} - \sqrt{|\mathcal{G}|})\|\mathbf{d}\|_2}{\sqrt{|\mathcal{B}||\mathcal{G}|}\Gamma_{|\mathcal{B}|} + |\mathcal{B}|\gamma_{|\mathcal{G}|}}\right] \tag{11}$$

is not empty and the system of inequalities

$$\begin{cases} \mu > \|\boldsymbol{\lambda}_{\mathcal{G}}\|_\infty \\ \alpha\|\mathbf{d}\|_2 > \|\boldsymbol{\lambda}_{\mathcal{G}}\|_1 + \mu|\mathcal{B}| \end{cases}$$

is satisfied with $\boldsymbol{\lambda}_{\mathcal{G}} = -\mathbf{J}_{\mathcal{G}}^{+}(\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}} + \mathbf{d})$ for every μ in the interval (11).

Proof The proof directly follows from Definitions 8 and 9, as well as Lemma 26 proved in Appendix C. ■

Lemma 11 provides an interval for the hyperparameter μ. The length of this interval and its location depend on the solution $\mathbf{x}$ and the amount of information in the measurements, but they are independent of the noise values. This is consistent with the existing results for the precedents of the quadratic regression problem, such as Lasso (Wainwright,

2009) and Graphical Lasso (Ravikumar et al., 2011). In such problems, the existence of this interval with unknown endpoints is enough for developing iterative methods, such as bisection techniques, to repeatedly solve the problem and update μ based on measuring the quality of estimation at each run of the optimization. This fits within the realm of model selection, where one can use information-theoretic methods such as the Akaike criterion. In Section 6, we will verify that the simple idea of trying multiple values for μ with different orders of magnitude performs well on real data.

4.2 Stochastic Bound

In the preceding section, we developed theoretical results on the correct recovery of the state of the problem and the number of permissible bad measurements. Unlike the existing results that focus on particular types of quadratic regression problems, these results apply to any arbitrary set of matrices $\mathbf{M}_r$'s. This generality of the results has made the conditions somewhat sophisticated. In what follows, we will simplify the results and provide some intuition under a stochastic setting.

Definition 12 *A matrix $\mathbf{X}$ is called standard Gaussian over $\mathbb{R}$ if its entries are independent and identically distributed random variables with a standard normal distribution.*

The data in this subsection is assumed to be stochastic, and therefore the associated theoretical results should be stated in a probabilistic sense. We select $\delta \in (0,1)$, and define $\varepsilon^* = \arg\min_{\varepsilon>0} 2\sqrt{6}e \cdot \frac{\sqrt{l \log \frac{3}{\varepsilon} + \log \frac{2}{\delta}}}{1-2\varepsilon}$ and $\tau_\delta = 2\sqrt{6}e \cdot \frac{\sqrt{l \log \frac{3}{\varepsilon^*} + \log \frac{2}{\delta}}}{1-2\varepsilon^*}$.

Theorem 13 *Consider a random instance of Problem (2) where the measurement matrices $\mathbf{M}_r$ and the exact solution $\mathbf{x}$ are random and distributed such that $\mathbf{J}$ (either $\bar{\mathbf{J}}$ or $\tilde{\mathbf{J}}$) is a standard Gaussian matrix. $\mathcal{B}$ consists of N elements selected uniformly on random from the measurement index set $\{1, 2, ..., m\}$. Consider an arbitrary constant $\delta \in (0,1)$. Introduce shortcut notation: $a = \sqrt{\sqrt{m-N} - \tau_\delta}$; $b = \sqrt{\sqrt{N} + \tau_\delta}$ and $c = \sqrt{\sqrt{m-N} + \tau_\delta}$, let $\bar{\alpha}$ be a constant satisfying*

$$\bar{\alpha} > \frac{(m-N)^{\frac{1}{4}} - N^{\frac{1}{4}}\frac{b}{a} + \frac{N}{a}\left[\frac{b}{N^{\frac{1}{4}}} + \frac{c}{(m-N)^{\frac{1}{4}}}\right]}{a - N^{\frac{1}{4}}(m-N)^{-\frac{1}{4}}b}$$

- *If the exact solution $\mathbf{x}$ of (2) and the prior knowledge $\hat{\mathbf{x}}$ are such that $\alpha_{SDP} \geq \bar{\alpha}$, then with probability at least $(1-\delta)^2$ there exists a constant μ for which $(\mathbf{x}\mathbf{x}^T, \boldsymbol{\eta})$ is the unique solution of the penalized SDP problem.*

- *If the exact solution $\mathbf{x}$ of (2) and the prior knowledge $\hat{\mathbf{x}}$ are such that $\alpha_{SOCP} \geq \bar{\alpha}$, then with probability at least $(1-\delta)^2$ there exists a constant μ for which $(\mathbf{x}\mathbf{x}^T, \boldsymbol{\eta})$ is the unique solution of the penalized SOCP problem.*

Proof The proof follows from Lemmas 1 and 6, together with Lemma 14 to be stated later in the paper. ■

Unlike the results of the previous subsection, the stochastic bounds given above are established for a "random scenario" when the adversary is oblivious and selects the indexes of the nonzero components of the error vector $\boldsymbol{\eta}$ on random with a uniform distribution. Nevertheless, we still consider the noise values to be completely arbitrary and possibly engineered to have the most negative impact on the regression problem.

It is important to discuss when $\mathbf{J}$ becomes a Gaussian matrix in order to use the stochastic bounds in Theorem 13. The easiest scenario corresponds to the case where the true solution $\mathbf{x}$ is a deterministic vector while $\mathbf{M}_r$'s are stochastic matrices. For example, $[\mathbf{M}_r]_{ij}$ with the distribution $\mathcal{N}(0, \frac{1}{nx_j^2})$ makes $[\mathbf{M}_r\mathbf{x}]$ a standard normal vector, independent of any other column vector in the matrix $\bar{\mathbf{J}}$. Likewise, an example of the data distribution for the SOCP case is $[\mathbf{M}_r]_{ii} \sim \mathcal{N}(0, n-1)$ and $[\mathbf{M}_r]_{ij} \sim \mathcal{N}(0, (\frac{x_i}{x_j})^2)$ when $i \neq j$. Indeed, $R_{ij} \sim \mathcal{N}(0,1)$ whenever $i \neq j$ will make $\tilde{\mathbf{J}}$ a standard Gaussian matrix.

The major difference between Theorem 13 and Theorem 10 is the elimination of the SSC property conditions. The simplification of the deterministic bounds was carried out for a Gaussian setting, but the developed techniques could be used to study other distributions as well. We will identify an interval for the hyperparameter μ below.

Lemma 14 *Let $\mathbf{J}$ be a matrix in $\mathbb{R}^{l\times m}$ that is sampled from a normal standard Gaussian distribution. Moreover, let $\mathbf{d}$ be a vector in $\mathbb{R}^l$ and $\boldsymbol{\lambda}$ be a vector in $\mathbb{R}^m$ such that $\boldsymbol{\lambda}_{\mathcal{B}} = \mu \cdot \mathbf{s}$, where μ is a scalar and the entries of $\mathbf{s}$ are only $+1$ or -1. Consider arbitrary numbers $\delta \in (0,1)$ and $\epsilon > 0$. Denote $\tau_{\delta,\varepsilon} = \frac{\sqrt{cl + c'\log\frac{2}{\delta}}}{1-2\varepsilon}$, where $c = 24e^2\log\frac{3}{\varepsilon}$ and $c' = 24e^2$. If*

$$\sqrt{|\mathcal{G}|} > \sqrt{|\mathcal{B}|}\frac{\sqrt{1+\Delta_{|\mathcal{B}|}}}{\sqrt{1-\Delta_{|\mathcal{G}|}}} + \frac{|\mathcal{B}|}{\alpha\sqrt{1-\Delta_{|\mathcal{G}|}}-1}\frac{\sqrt{1+\Delta_{|\mathcal{B}|}}+\sqrt{1+\Delta_{|\mathcal{G}|}}}{\sqrt{1-\Delta_{|\mathcal{G}|}}}, \tag{12}$$

where $\Delta_t \geq \frac{\tau_{\delta,\varepsilon}}{\sqrt{t}}$ for $t = |\mathcal{B}|$ and $|\mathcal{G}|$, then with probability at least $(1-\delta)^2$ the interval

$$\left[\frac{\|\mathbf{d}\|_2}{\sqrt{|\mathcal{G}|(1-\Delta_{|\mathcal{G}|})}-\sqrt{|\mathcal{B}|(1+\Delta_{|\mathcal{B}|})}}, \frac{(\alpha\sqrt{(1-\Delta_{|\mathcal{G}|})}-1)\|\mathbf{d}\|_2}{|\mathcal{B}|(\sqrt{(1+\Delta_{|\mathcal{B}|})}+\sqrt{(1+\Delta_{|\mathcal{G}|})})}\right] \tag{13}$$

is not empty and the system of inequalities

$$\begin{cases} \mu > \|\boldsymbol{\lambda}_{\mathcal{G}}\|_\infty \\ \alpha\|\mathbf{d}\|_2 > \|\boldsymbol{\lambda}_{\mathcal{G}}\|_1 + \mu|\mathcal{B}| \end{cases}$$

is satisfied with $\boldsymbol{\lambda}_{\mathcal{G}} = -\mathbf{J}_{\mathcal{G}}^{+}(\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}} + \mathbf{d})$ for every μ in the interval (13).

Proof The proof is provided in Appendix C ■

As explained after Lemma 11, the existence of the unknown interval given in (13) enables the design of iterative techniques to adaptively find a suitable value for μ. We will illustrate the insensitivity of the conic programs to the exact value of μ in Section 6.

It can be verified that to satisfy the condition (12), the number of good measurements $|\mathcal{G}|$ must grow quadratically in the number of bad measurements $|\mathcal{B}|$ for both the penalized

SDP and the penalized SOCP relaxations. Nevertheless, in the case when $\alpha \to \infty$, this condition on $|\mathcal{G}|$ and $|\mathcal{B}|$ can be reduced to the simple inequality

$$|\mathcal{G}|(1 - \frac{\tau_{\delta,\varepsilon}}{\sqrt{|\mathcal{G}|}}) > |\mathcal{B}|(1 + \frac{\tau_{\delta,\varepsilon}}{\sqrt{|\mathcal{B}|}})$$

or equivalently

$$\begin{cases} |\mathcal{G}| > \tau_{\delta,\varepsilon}^2 \\ |\mathcal{B}| < |\mathcal{G}| + \tau_{\delta,\varepsilon}^2 - 2\tau_{\delta,\varepsilon}\sqrt{|\mathcal{G}|} \end{cases} \tag{14}$$

The above inequalities imply that the number of bad measurements $|\mathcal{B}|$ is allowed to increase from $\mathcal{O}(\sqrt{|\mathcal{G}|})$ to $\mathcal{O}(|\mathcal{G}|)$ as the amount of information in the prior knowledge increases.

Numerical studies show that the function $\tau_{\delta,\varepsilon}$ is expected to be fairly flat with respect to ε for practically important values of the parameters. For illustration purposes, consider $l = 100$ and $\delta = 0.05$. In this setting, $\varepsilon = \varepsilon^* = 0.05514$ is the minimum of $\tau_{\delta,\varepsilon}$ and

$$\tau_{\delta,\varepsilon^*}^2 \simeq 893.7l + 223.6 \log \frac{2}{\delta} \simeq \tau_\delta^2$$

which demonstrates the asymptotic behavior of the function. Since $l_{SDP} = n$ but $l_{SOCP} = n^2 - n$, it can be concluded that the guarantee for the SDP approach works whenever the number of measurements is on order of the size of the problem, while the guarantee for the SOCP approach requires a higher number of measurements. This gives rise to a salient difference between the penalized SDP and the penalized SOCP: the SDP approach offers a higher performance over the SOCP approach but is computationally more expensive. Another important difference between the SDP and SOCP approaches—coming from the nature of the problem itself—is rooted in the definition of the coefficient α. The amount of prior knowledge needed for the SDP approach to work is less than or equal to that needed for the SOCP approach.

5. Robust Least-Squares Regression

Taking a step back from the penalized convex program, note that the literature on regression under sparse noise utilizes other methods along with convex relaxation techniques. To consider an alternative baseline, in this section we focus our attention on the development of an iterative technique inspired by Bhatia et al. (2017). This new method is most useful when no prior knowledge about the unknown solution is available. To build an iterative algorithm for solving Problem (2), consider the optimization

$$\begin{array}{ll} \underset{\mathbf{W}\in\mathbb{S}^n,\ \nu\in\mathbb{R}^m}{\text{minimize}} & \dfrac{1}{2}\displaystyle\sum_{r=1}^{m} \left(\langle \mathbf{W}, \mathbf{M}_r\rangle + \nu_r - y_r\right)^2 \\ \text{subject to} & \mathbf{W} \succeq_{\mathcal{C}} 0 \\ & \|\boldsymbol{\nu}\|_0 \le k \end{array} \tag{15}$$

where k is a parameter. This problem is nonconvex due to a cardinality constraint.

Definition 15 *Define $HT_k(\mathbf{y}) : \mathbb{R}^m \to \mathbb{R}^m$ as a hard thresholding operator such that*

$$[HT_k(\mathbf{z})]_i = \begin{cases} z_i & \text{if } |z_i| \text{ is among the } k \text{ largest-in-magnitude entries of } \mathbf{z} \\ 0 & \text{otherwise,} \end{cases}$$

where $[HT_k(\mathbf{z})]_i$ denotes the i^{th} entry of $HT_k(\mathbf{z})$.

Consider the function

$$f(\boldsymbol{\nu}) := \min_{\mathbf{W} \succeq_{\mathcal{C}} 0} \frac{1}{2} \sum_{r=1}^{m} \left(\langle \mathbf{W}, \mathbf{M}_r \rangle - (y_r - \nu_r) \right)^2$$

and let $\hat{\mathbf{W}}(\boldsymbol{\nu})$ denote a solution to this problem. We propose a Hard Thresholding method for solving the quadratic regression problem, which consists of the iterative scheme

$$\boldsymbol{\nu}^{t+1} = HT_k(\boldsymbol{\nu}^t - \mathbf{d}(\boldsymbol{\nu}^t))$$

where

$$\mathbf{d}(\boldsymbol{\nu}) = \frac{1}{2} \nabla_{\boldsymbol{\nu}} \left(\sum_{r=1}^{m} \left(\langle \mathbf{W}, \mathbf{M}_r \rangle - (y_r - \nu_r) \right)^2 \right) \Bigg|_{\mathbf{W} = \hat{\mathbf{W}}(\boldsymbol{\nu})}$$

(the symbol $\nabla_{\boldsymbol{\nu}}$ stands for the gradient with respect to $\boldsymbol{\nu}$). By Lemma 3.3.1 in Bertsekas (1995), if $\hat{\mathbf{W}}(\boldsymbol{\nu})$ is a continuously differentiable mapping, then $\nabla f(\boldsymbol{\nu}) = \mathbf{d}(\boldsymbol{\nu})$. Inspired by this fact, one may informally regard $\mathbf{d}(\boldsymbol{\nu})$ as the gradient of the optimal value of the optimization problem (15) without its cardinality constraint. Define $\mathbf{w} = \text{vec}(\mathbf{W})$, $\hat{\mathbf{w}}(\boldsymbol{\nu}) = \text{vec}(\hat{\mathbf{W}}(\boldsymbol{\nu}))$, $\mathbf{a}_r = \text{vec}(\mathbf{M}_r)$ for $r = 1, ..., m$, and $\mathbf{A} = [\mathbf{a}_1 \ \dots \ \mathbf{a}_m]^T$. It can be verified that

$$\mathbf{d}(\boldsymbol{\nu}) = \mathbf{A}\hat{\mathbf{w}}(\boldsymbol{\nu}) - \mathbf{y} + \boldsymbol{\nu}$$

which implies that

$$HT_k(\boldsymbol{\nu} - \mathbf{d}(\boldsymbol{\nu})) = HT_k(\mathbf{y} - \mathbf{A} \cdot \text{vec}(\hat{\mathbf{W}}(\boldsymbol{\nu})))$$

Based on this formula, we propose a conic hard thresholding method in Algorithm 1. Unlike

Algorithm 1 Conic Hard Thresholding

Input: Covariates $\mathbf{A}$, responses $\mathbf{y}$, corruption index k, tolerance ε, and cone $\mathcal{C}$

Initialization :

$\boldsymbol{\nu}^0 \leftarrow \mathbf{0}$, $t \leftarrow 0$;

LOOP Process

while $\|\boldsymbol{\nu}^t - \boldsymbol{\nu}^{t-1}\| > \varepsilon$ **do**

  $\hat{\mathbf{W}}^t = \arg \min_{\mathbf{W} \succeq_{\mathcal{C}} \mathbf{0}} \sum_{r=1}^{m} \left(\langle \mathbf{W}, \mathbf{M}_r \rangle - (y_r - \nu_r^t) \right)^2$;

  $\boldsymbol{\nu}^{t+1} = HT_k(\mathbf{y} - \mathbf{A} \cdot \text{vec}(\hat{\mathbf{W}}^t))$;

  $t \leftarrow t + 1$;

end while

return $\hat{\mathbf{W}}^{t+1}$

the penalized SDP and penalized SOCP methods, Algorithm 1 does not rely on any prior knowledge. Instead of the penalty terms in the objective, it solves a sequence of conic programs to identify the set of bad measurements through a thresholding technique. In the regime where $m \geq \frac{n(n+1)}{2}$, this algorithm with a high computational complexity can be further relaxed by letting the cone $\mathcal{C}$ be the set of symmetric matrices. We refer to this as **Algorithm 2**, where the condition $\mathbf{W} \succeq_{\mathcal{C}} \mathbf{0}$ is reduced to $\mathbf{W} = \mathbf{W}^T$. Note that Algorithm

2 is not effective if $m < n(n+1)/2$ because the number of measurements becomes less than the number of scalar variables in $\mathbf{W}$. On the other hand, as m grows, the feasibility constraint $\mathbf{W} \succeq_{\mathcal{C}} \mathbf{0}$ would more likely be satisfied for free (since the feasible set shrinks) and Algorithm 1 would perform similarly to Algorithm 2. Inspired by this property, we analyze the asymptotic behavior of Algorithm 2 for Gaussian systems below.

Lemma 16 *Suppose that* $|\mathcal{B}| < \frac{m}{20000}$, $m \geq n^2$, *and* $\mathbf{M}_r$ *is a random normal Gaussian matrix for* $r = 1, ..., m$. *For every* $\epsilon > 0$, *Algorithm 2 recovers a matrix* $\mathbf{W}$ *such that* $\|\mathbf{W} - \mathbf{x}\mathbf{x}^T\|_2 \leq \epsilon$ *within* $\mathcal{O}(\log(\frac{\|\boldsymbol{\eta}\|_2}{\epsilon}) + \log(\frac{2m}{n^2+n}))$ *iterations.*

Proof It follows from Theorem 4 by Bhatia et al. (2017). ■

Let $\mathbf{W}^*$ be any solution obtained by Algorithm 2. Then, one can use its eigenvalue decomposition to find a vector $\mathbf{u}$ such that $\mathbf{u} = \arg\min_{\mathbf{v} \in \mathbb{C}^n} \|\mathbf{v}\mathbf{v}^T - \mathbf{W}\|_2$. Therefore,

$$\begin{aligned} \|\mathbf{u}\mathbf{u}^T - \mathbf{x}\mathbf{x}^T\|_2 &= \|(\mathbf{u}\mathbf{u}^T - \mathbf{W}^*) - (\mathbf{x}\mathbf{x}^T - \mathbf{W}^*)\|_2 \\ &\leq \|\mathbf{u}\mathbf{u}^T - \mathbf{W}^*\|_2 + \|\mathbf{x}\mathbf{x}^T - \mathbf{W}^*\|_2 \leq 2\varepsilon \end{aligned} \tag{16}$$

This means that Algorithm 2 can be used to find an approximate solution $\mathbf{u}$ with any arbitrary precision for the robust regression problem for Gaussian systems with a large number of measurements and yet it allows up to a constant fraction of measurements to be completely wrong. Comparing this with the guarantee $O(|\mathcal{B}|) = O(|\mathcal{G}|^{\frac{1}{2}})$ for the penalized conic methods, given by Theorem 13, it can be concluded that Algorithm 1 (or 2) is asymptotically more robust to outliers than the penalized conic program since it solves a sequence of optimization problems iteratively as opposed to a single one. This leads to another level of tradeoff between the complexity of an estimation method and its robustness level.

The theoretical analyses of this work were all on a regression model subject to a sparse error vector. However, the results can be slightly modified to account for modest noise values in addition to sparse errors. The bounds derived in this work remain the same, but the solutions found by the penalized conic problem and Algorithm 1 would no longer match the true regression solution being sought (as expected, due to a corruption in all equations). The mismatch error is a function of the modest noise values. The details are omitted for brevity; however, the this scenario will later be analyzed in numerical examples.

6. Experiments

In this section, we study the numerical properties of the penalized conic methods and the conic hard thresholding Algorithm 1. The simulation results in Section 6.2 and 6.3 are on physical systems for which we use the realistic prior knowledge that can be inferred from the physics of the problem (without having access to any information about the solution). In addition, we do not use any prior knowledge for the simulations in Sections 6.1 and 6.4, and select $\mathbf{M}$ to simply be a diagonal matrix.

6.1 Synthetic Data

Following Madani et al. (2017a), we define the sparsity pattern of an arbitrary matrix $\mathbf{X} \in \mathbb{S}^n$ to be a binary matrix $\mathbf{N} \in \mathbb{S}^n$ whose (i,j)-entry is equal to 1 if and only if $X_{ij} \neq 0$.

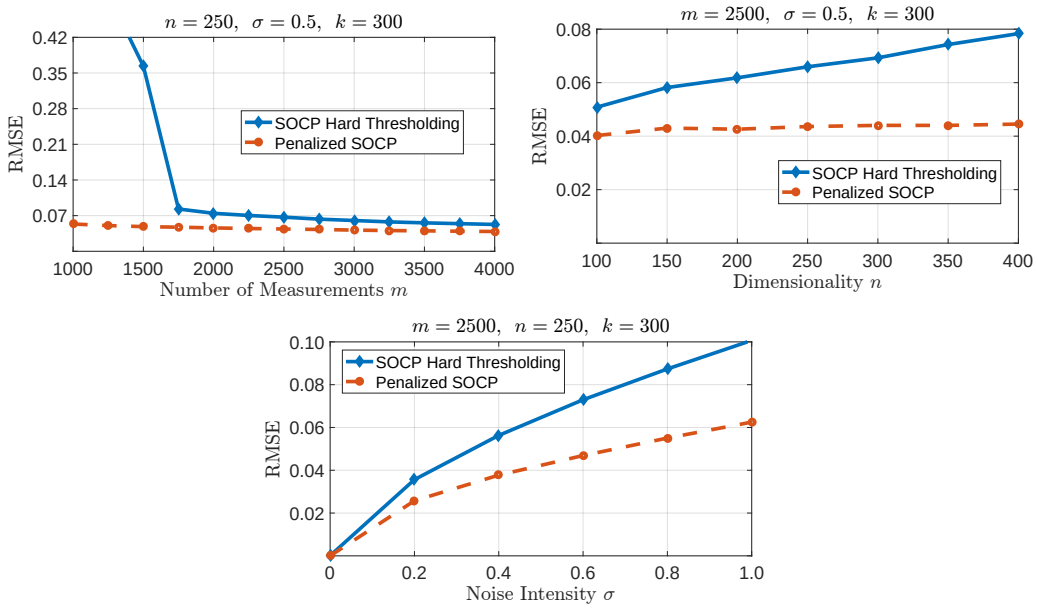

Figure 1: Estimation error as a function of: (a) the number of data points m, (b) the dimensionality n, (c) the standard deviation σ of additive white noise.

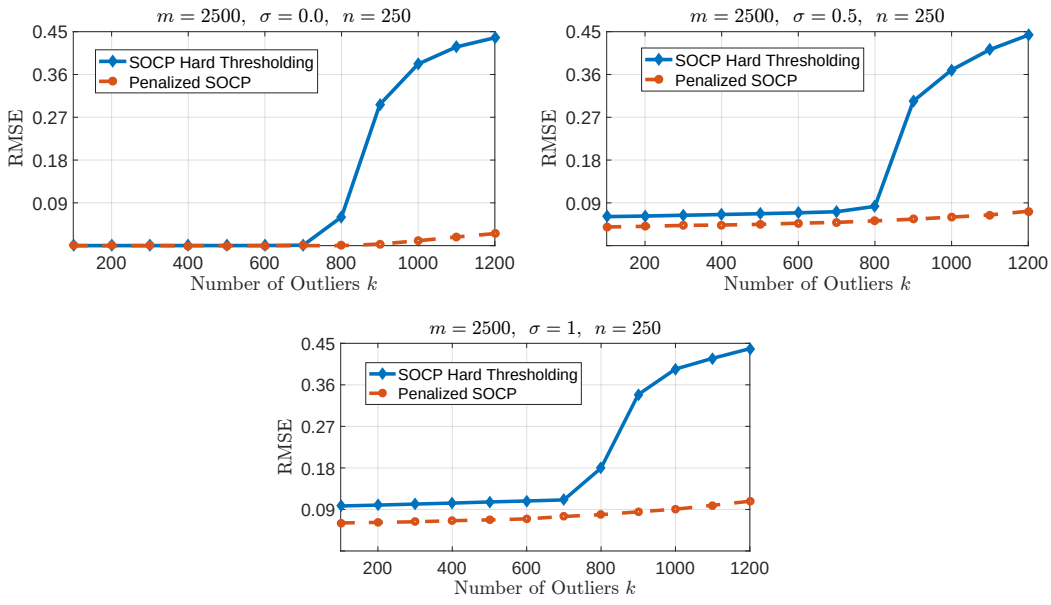

Figure 2: Estimation error as a function of the number of bad measurements k for different magnitudes of additive dense Gaussian noise.

Define the set

$$\mathcal{S}(\mathbf{N}) \triangleq \{\mathbf{X} \in \mathbb{S}^n | \mathbf{X} \circ \mathbf{N} = \mathbf{X}\}$$

We conduct some experiments on synthetically generated quadratic regression data sets with corruptions. The true model vector $\mathbf{x}$ is chosen to be a random unit-norm vector, while the input matrices $\mathbf{M}_r$'s are chosen from $\mathcal{S}(\mathbf{N})$ according to a common random sparsity pattern $\mathbf{N}$. The nonzero entries of $\mathbf{M}_r$'s are generated from a normal standard distribution. The matrix $\mathbf{N}$ has all diagonal elements and $3n$ off-diagonal elements nonzero. The off-diagonal positions are selected uniformly. The measurements to be corrupted are chosen uniformly at random and the value of each corruption is generated uniformly from the interval $[10, 20]$. The measurements are then generated as $y_r = \mathbf{x}^* \mathbf{M}_r \mathbf{x} + \eta_r + \omega_r$, where in addition to the sparse error vector $\boldsymbol{\eta}$ there is a random dense noise vector $\boldsymbol{\omega}$ whose entries are Gaussian with zero mean and standard deviation σ. All reported results are averaged over 10 random trials.

By assuming that no prior information about the solution $\mathbf{x}$ is available, we set the matrix $\mathbf{M}$ to be equal to $\mathbf{I}_n$. The parameter μ is chosen as 10^{-2}. Regarding Algorithm 1, the parameter k is selected as the true number of corrupted measurements, the tolerance ε is set to 10^{-3}, and the algorithm is terminated early if the number of conic iterations exceeds 50. In both of the methods, $\mathcal{C}$ is considered to be the $2\mathcal{PSM}$ cone. Hence, we refer to these methods as penalized SOCP and SOCP hard thresholding. Due to the sparsity in the data, the SOCP formulation can be simplified by only imposing those 2×2 constraints in (2) that correspond to the members of $\{(i, j) \mid N_{ij} = 1\}$.

We measure the performance of each algorithm using the root-mean-squared error (RMSE) defined as $\frac{\|\mathbf{x}^\star - \mathbf{x}\|_2}{\sqrt{n}}$, where x is the output of the algorithm and $x^\star$ is the correct solution. Figure 1 shows the RSME in three different plots as a function of the number of data points m, the dimensionality n, and the additive white noise standard deviation σ. Figure 2 depicts the RSME as a function of the number of bad measurements k for different magnitudes of additive dense Gaussian noise. It can be observed that both the penalized conic problem and the conic hard thresholding algorithm exhibit an exact recovery property for systems with up to 700 randomly corrupted measurements out of 2500 measurements in the absence of dense Gaussian noise. The same behavior is observed in the presence of dense Gaussian noise of different magnitudes: the error of the penalized SOCP solution grows gradually, while the error of the hard thresholding algorithm has a jump at around 800 bad measurements. These simulations support the statement that up to a constant fraction of measurements could be completely wrong, and yet the unknown regression solution is found precisely.

Although the theoretical analysis provided in this paper favors Algorithm 1 over the penalized conic problem, our empirical analysis shows that the penalized SOCP method has a better performance than the hard thresholding algorithm uniformly in the number of measurements, dimensionality, noise magnitude and the number of outliers. To explain this observation, note that the derived theoretical bounds correspond to the worst-case scenario and are more conservative for an average scenario. Moreover, the implementation of Algorithm 1 in this section has limited the number of iterations to 50, while Theorem 16 requires the number of iterations to grow with respect to the amount of corruption.

The results of this part are produced using the standard MOSEK v7. SOCP-solving procedure, run in MATLAB on a 12-core 2.2GHz machine with 256GB RAM. The CPU

time for each round of solving SOCP ranges from 3s (for $n = 250$, $m = 2500$) to 30s (for $n = 400$, $m = 2500$).

6.2 State Estimation for Power Systems

In this subsection, we present empirical results for the penalized conic problem with a PSD cone $\mathcal{C}$ tested on the real data for the power flow state estimation with outliers. As discussed in Madani et al. (2017b), this problem can be formulated as robust quadratic regression. The experiment is run on the PEGASE 1354-bus European system borrowed from the MATPOWER package (Fliscounakis et al., 2013; Josz et al., 2016). This system has 1354 nodes and the objective is to estimate the nodal voltages based on voltage magnitude and power measurements of the form $y_r = \mathbf{x}^*\mathbf{M}_r\mathbf{x} + \eta_r + \omega_r$, where $\boldsymbol{\omega}$ is a dense additive noise whose r^{th} entry is Gaussian with mean zero and the standard deviation equal to σ times the true value of the corresponding voltage/power parameter. The dimension of the complex vector $\mathbf{x}$ is 1354, which leads to 2708 real variables in the problem. In this model, the measurements are voltage magnitude squares, active and reactive nodal power injections, and active and reactive power flows from both sides of every line of the power system. This amounts to $3n + 4t = 12026$ measurements, where $t = 1991$ denotes the number of lines in the system. Note that the quadratic regression problem is complex-valued in this case.

The penalty parameter μ of the penalized conic problem is set to 10^2 and the matrix $\mathbf{M}$ is chosen as $-\mathbf{Y}+\gamma\mathbf{I}$, where $\mathbf{Y}$ is the susceptance matrix of the system and γ is the smallest positive number that makes $\mathbf{M}$ positive semidefinite. This choice of $\mathbf{M}$ corresponds to $\hat{\mathbf{x}}$ being equal the eigenvector of $-\mathbf{Y}$ associated with its smallest eigenvalue. This eigenvector provides a combination of voltages that results in the minimum amount of reactive power loss, as shown by Madani et al. (2019). Hence, by using this particular $\mathbf{M}$, we implicitly assume that the ground truth vector of voltages does not create a large amount of reactive power loss, which is a physical feature of real-world power systems. Since the penalized SDP problem is large-scale, we employ a tree decomposition technique to leverage the sparsity of the problem to solve it more efficiently (Madani et al., 2016). The width of the tree decomposition used to reduce the complexity is equal to 12. We do not report any results on Algorithm 1 because it requires solving large-scale SDPs successively and this could be time-consuming. Moreover, the number of measurements is not high enough to use Algorithm 2, and, therefore, we will not test this method either.

The numerical results are reported in Figure 3. Remarkably, if the dense Gaussian noise is non-existent, the conic problem recovers the solution precisely as long as the number of bad measurements is less than 150 (note that $\sqrt{m} \simeq 109$). Note that power systems are sparse networks, their models are far from Gaussian, but the bounds from Theorem 13 are still valid in this numerical example.

6.3 Dynamic State Estimation

In this subsection, we demonstrate the usefulness of the proposed mathematical technique for solving sequential decision-making problems. We again consider data analytics for power systems, but leverage the fact that state estimation is solved regularly due to the time-varying and stochastic demands requested by millions of consumers. At each time instance, we use the estimated state of the system at the previous time as prior knowledge for infer-

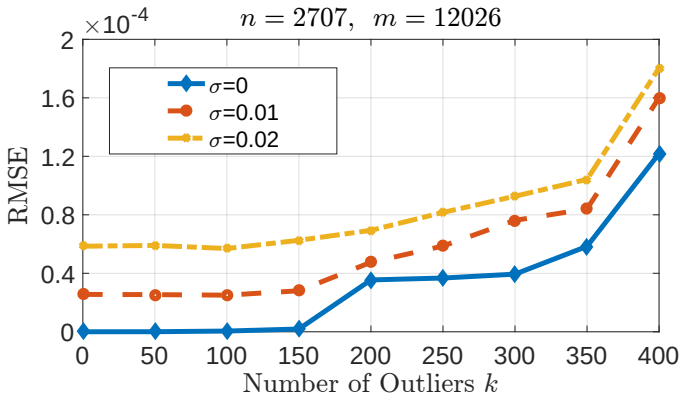

Figure 3: This plot shows the RMSE with respect to the number of corrupted measurements k for the PEGASE 1354-bus system.

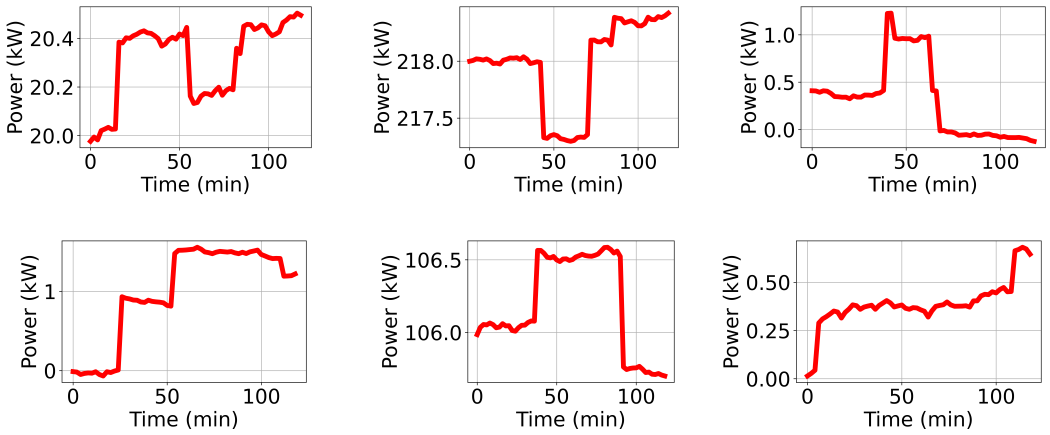

Figure 4: Net active (top) and reactive (bottom) powers at buses 2 (left), 50 (middle) and 93 (right) over the period of simulation.

ring the current state of the system. We use the IEEE 300-bus benchmark system from the MATPOWER package and simulate two hours of its evolution under varying nodal active and reactive powers to reflect the changes in supply and demand. The net nodal powers are the only time-varying measurements of this system (each net power is the difference between the generation and the consumption at the node). To make the analysis realistic, we simulate both continuous changes and sudden jumps in the time-varying nodal powers. The continuous changes are modeled by a Wiener random process, while the jump values and locations are sampled from a uniform distribution that affect each time-varying measurement (curve) 5 times over the considered interval on average. The evolution of some of these measurements is depicted in Fig. 4.

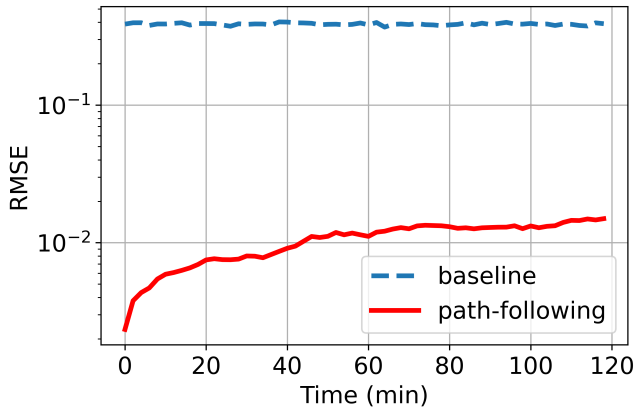

Figure 5: This plot shows the root mean squared error over the time period of the simulation. The dashed line denotes the error obtained by applying the SOCP penalized method with the objective matrix $\mathbf{M}$ constructed from the matrix $\mathbf{Y}$ as in Section 6.2. The solid line denotes the error obtained by applying the same method, but using a dynamic method for designing $\mathbf{M}$ through the path-following approach.

At each time step of the simulation, happening every 2 minutes, we solve the state estimation under sparse noise via the penalized SOCP method described in Section 3.2. We let the number of corrupted measurements be 20% of the total number of measurements $m = 3444$. In the first time step, we construct the matrix $\mathbf{M}$ according to the formula (7) based on the true state of the system, set in accordance with the IEEE 300-bus system data set. At each subsequent time instance, we construct the matrix $\mathbf{M}$ based on the solution obtained in the previous time step. We refer to this procedure as the path-following experiment.

As a baseline for comparison, we also study a different strategy where we apply the penalized SOCP method at each time step with the objective matrix $\mathbf{M}$ constructed from the matrix $\mathbf{Y}$ as in Section 6.2, without using the prior knowledge in the solution of the previous time instance. Both the baseline and path-following experiments were conducted 5 times to produce an average result. The values of the parameter μ were chosen prior to the experiment as $5 \cdot 10^{-1}$ for the baseline and $5 \cdot 10^{-4}$ for the path-following part. They were chosen experimentally from the set $\{5, 5 \cdot 10^{-1}, 5 \cdot 10^{-2}, 5 \cdot 10^{-3}, 5 \cdot 10^{-4}, 5 \cdot 10^{-5}\}$.

Figure 5 demonstrates that the errors produced in the path-following experiment are smaller than the errors produced in the baseline experiment by an order of magnitude. Given that the only difference between these two approaches is in the construction of the objective matrix, one can conclude that the solution of the problem in each time step can serve as useful prior knowledge for the next time step.

There is also a notable uptrend of the bottom curve, which reflects the error accumulation during the path-following experiment. This is due to the fact that the prior knowledge at each time is considered to be the state estimated at the previous time, rather than the true state of the system at the previous time. As a result, if the previous estimated state has some error, it affects learning the current state and this error accumulates over time.

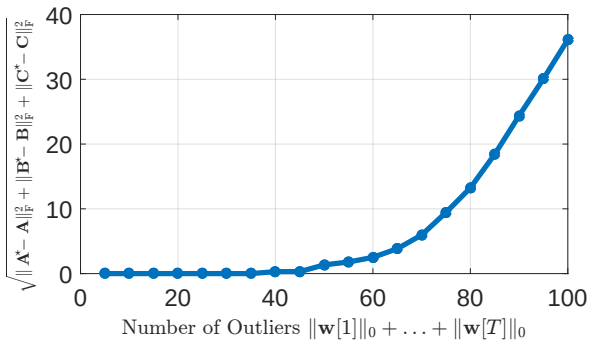

Figure 6: This plot shows the average estimation error of 15 random ground truth realizations with respect to the number of corrupted observations.

However, since the model (e.g., topology) of a power network changes on a slow time scale (e.g., every few hours), there is a reset in the process that eliminates the error.

6.4 Linear System Identification

Following Fattahi and Sojoudi (2018), this case study is concerned with the problem of identifying the parameters of a linear dynamical system, given limited observation and non-uniform snapshots of the state vector. Consider a discrete-time linear system described by the equations

$$\mathbf{x}^\star[\tau+1] = \mathbf{A}^\star\mathbf{x}^\star[\tau] + \mathbf{B}^\star\mathbf{u}[\tau] \qquad \tau = 1, 2, \ldots, T-1, \tag{17a}$$

$$\mathbf{y}[\tau] = \mathbf{C}^\star\mathbf{x}^\star[\tau] + \mathbf{w}^\star[\tau] \qquad \tau = 1, 2, \ldots, T, \tag{17b}$$

where

- $\{\mathbf{x}^\star[\tau] \in \mathbb{R}^n\}_{\tau=1}^T$ are the state vectors that are known at times $\tau \in \{\tau_1, \ldots, \tau_o\}$,
- $\{\mathbf{u}[\tau] \in \mathbb{R}^m\}_{\tau=1}^T$ and $\{\mathbf{y}[\tau] \in \mathbb{R}^k\}_{\tau=1}^T$ are the known control and observation vectors, respectively,
- $\mathbf{A}^\star \in \mathbb{R}^{n\times n}$, $\mathbf{B}^\star \in \mathbb{R}^{n\times m}$ and $\mathbf{C}^\star \in \mathbb{R}^{k\times n}$ are fixed unknown matrices, and
- $\{\mathbf{w}^\star[\tau] \in \mathbb{R}^k\}_{\tau=1}^T$ are the vectors of sparsely occurring observation errors that are unknown.

We propose to determine the triplet $(\mathbf{A}^\star, \mathbf{B}^\star, \mathbf{C}^\star)$ by solving the following system of quadratic equations:

$$0 = e \times \mathbf{x}[\tau+1] - (e \times \mathbf{B})\mathbf{u}[\tau] - \mathbf{A}\mathbf{x}[\tau] \qquad \tau = 1, 2, \ldots, T-1, \tag{18a}$$

$$\mathbf{y}[\tau] = \mathbf{C}\mathbf{x}[\tau] + \mathbf{w}[\tau] \qquad \tau = 1, 2, \ldots, T, \tag{18b}$$

$$\mathbf{x}[\tau] = e \times \mathbf{x}[\tau] \qquad \tau = \tau_1, \tau_2, \ldots, \tau_o, \tag{18c}$$

$$1 = e^2, \tag{18d}$$

with the unknown vector

$$\mathbf{z} \triangleq [e,\ \mathbf{x}[1]^\top,\ \mathbf{x}[2]^\top,\ \ldots,\ \mathbf{x}[T]^\top, \mathrm{vec}\{\mathbf{A}\}^\top, \mathrm{vec}\{\mathbf{B}\}^\top, \mathrm{vec}\{\mathbf{C}\}^\top]^\top \tag{19}$$

and the noise estimation vectors $\{\mathbf{w}[\tau] \in \mathbb{R}^k\}_{\tau=1}^T$. The auxiliary variable e is added to make the system of equations homogeneous, similar to the canonical quadratic regression problem (2). In order to solve the system of equations (18), we formulate the penalized SDP problem (3) by introducing the matrix variable $\mathbf{Z}$ accounting for $\mathbf{z}\mathbf{z}^\top$. In this experiment, we use the objective function

$$\langle \mathbf{M}, \mathbf{Z} \rangle + \eta \times \sum_{\tau=1}^{T} \|\mathbf{w}[\tau]\|_1 \tag{20}$$

where $\mathbf{M} = \mathrm{diag}\{[0,\ 0.001 \times \mathbf{1}_{1\times nT},\ \mathbf{1}_{1\times n^2},\ \mathbf{0}_{1\times nm},\ \mathbf{1}_{1\times nk}]\}$ and $\eta = 0.1$.

We consider system identification problems with $n = 5$, $m = 2$, $k = 3$, and $T = 50$ time epochs. We assume that, for every $\tau \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, the state vector $\mathbf{x}^\star[\tau]$ is unknown. The elements of the ground truth matrices $\mathbf{A}^\star \in \mathbb{R}^{5\times 5}$, $\mathbf{B}^\star \in \mathbb{R}^{5\times 2}$, $\mathbf{C}^\star \in \mathbb{R}^{3\times 5}$ and the control vectors $\{\mathbf{u}[\tau]\}_{\tau=1}^T$, as well as the initial state $\mathbf{x}^\star[1]$ have independent Gaussian distribution with zero mean and variance $\frac{1}{3}$. Unstable ground truth matrices $\mathbf{A}$ with an eigenvalue outside of the unit circle are excluded. For various values of ρ, we randomly choose ρ elements of $\{\mathbf{y}[\tau]\}_{\tau=1}^T$ and corrupt them by adding observation errors chosen uniformly from the interval $[10, 20]$. Figure 6, demonstrates the average estimation error for 15 trials. As shown in Figure 6, with up to 35 corrupted observations, the triplet $(\mathbf{A}^\star, \mathbf{B}^\star, \mathbf{C}^\star)$ can be recovered with zero error. Exploiting the sparsity of the problems (Nakata et al., 2003), each round of penalized SDP has been solved within 5 minutes.

7. Conclusion

In this paper, we study instances of the $\mathcal{NP}$-hard problem of learning a parametric model from a series of nonlinear data points subject to sparse adversarial errors of arbitrary magnitudes. This problem arises in the data analysis process of cyber-physical systems, and so it is of both theoretical and practical importance to construct polynomial algorithms for this problem. We develop two conic programming methods to learn the unknown model, both of which accepts any available prior knowledge about the solution. Sufficient conditions are derived to guarantee the success of these methods, and it is shown that the trade-off between the developed techniques is in their performance versus required computational power. In the case when no prior knowledge is available, a surrogate iterative method based on conic programming is developed, offering another level of this trade-off. These methods are studied under a stochastic setting, and it is shown that they can correctly find the model even if up to a constant factor of measurements are strategically corrupted. The results are demonstrated in four experiments on learning dynamical systems and power network states.

Acknowledgments

This work was supported by grants from AFOSR, DARPA, ONR, ARO and NSF.

Appendix A.

The following lemma studies Slater's condition for the dual problem (4).

Lemma 17 *If there exists an index $r \in \{1, \dots, m\}$ such that $\hat{\mathbf{x}}^T \mathbf{M}_r \hat{\mathbf{x}} \neq 0$, then the interior of the feasible region of the problem (4) is not empty and strong duality holds for the penalized SDP.*

Proof Choose $c \in \{-1, +1\}$ such that $c\hat{\mathbf{x}}^T \mathbf{M}_r \hat{\mathbf{x}} > 0$. To construct a strictly feasible point for Problem (4), it is enough to consider $\boldsymbol{\lambda} = u\mathbf{e}_r$, where $u > 0$ is a constant that is smaller than μ and $\mathbf{M} + cu\mathbf{M}_{\hat{r}} \succ 0$. Such a constant exists due to Lemma 3.2.1 in Bertsekas (1999). ■

Proof of Lemma 1 Strong duality of the penalized SDP follows from Lemma 17. We aim to prove that under such a choice of $\hat{\boldsymbol{\lambda}}$, the matrix $\mathbf{M} + \sum \hat{\lambda}_r \mathbf{M}_r$ is a PSD matrix. The complementary slackness condition:

$$\langle \mathbf{x}\mathbf{x}^T, \mathbf{M} + \sum_{r=1}^{m} \lambda_r \mathbf{M}_r \rangle = 0$$

or equivalently

$$(\mathbf{M} + \sum_{r=1}^{m} \lambda_r \mathbf{M}_r)\mathbf{x} = \mathbf{0}. \tag{21}$$

It is straightforward to verify that the condition (21) is satisfied for $\lambda = \hat{\boldsymbol{\lambda}}$. Therefore, $\operatorname{rank}(\mathbf{M} + \sum_{r=1}^{m} \hat{\boldsymbol{\lambda}}_r \mathbf{M}_r) \leq n - 1$. In light of Corollary 4.3.39 in Horn (2013), that $\kappa(\cdot)$ is a concave function. Now, it follows from condition (6b) that

$$\kappa(\mathbf{M} + \sum_{r=1}^{m} \hat{\lambda}_r \mathbf{M}_r) \geq \kappa(\mathbf{M}) + \sum_{r=1}^{m} \kappa(\hat{\lambda}_r \mathbf{M}_r) \geq \kappa(\mathbf{M}) - 2\sum_{r=1}^{m} |\hat{\lambda}_r| \ \|\mathbf{M}_r\|_2 \quad > 0$$

which, combined with (6b), yields that $\operatorname{rank}(\mathbf{M} + \sum_{r=1}^{m} \hat{\lambda}_r \mathbf{M}_r) \geq n - 1$. Dual feasibility for $\hat{\boldsymbol{\lambda}}_{\mathcal{G}}$ follows from condition (21), the above inequality, definition of κ and condition (6a). On the other hand, primal feasibility is satisfied for $(\mathbf{x}\mathbf{x}^T, \eta)$. Therefore, $(\mathbf{x}\mathbf{x}^T, \eta)$ and $\hat{\boldsymbol{\lambda}}$ is a primal-dual optimal pair for the problem. This completes the proof. ■

Appendix B.

Lemma 18 *The sequence $\{\mathbf{A}^{ij} \in \mathbb{S}^2\}_{i<j}$ is a decomposition of $\mathbf{A}$ if and only if:*

$$\begin{cases} \left[\mathbf{A}^{ij}\right]_{21} = \left[\mathbf{A}^{ij}\right]_{12} = A_{ij} = A_{ji} \\ \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left[\mathbf{A}^{ji}\right]_{22} + \sum_{j=2}^{n} \sum_{i=1}^{j-1} \left[\mathbf{A}^{ij}\right]_{11} = \mathbf{A}_{ii} \end{cases}$$

Proof The proof is based on basis linear algebra and is omitted for brevity. ■

We define linear operations over decompositions below.

Definition 19 *Given the sequences* $\{\mathbf{A}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and* $\{\mathbf{B}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$, *define the sum:*

$$\{\mathbf{A}^{ij}\}_{i<j}^{j\leq n} + \{\mathbf{B}^{ij}\}_{i<j}^{j\leq n} := \{\mathbf{A}^{ij} + \mathbf{B}^{ij}\}_{i<j}^{j\leq n}$$

Definition 20 *For a sequence* $\{\mathbf{A}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and a scalar* $c \in \mathbb{R}$, *define the multiplication:*

$$c\{\mathbf{A}^{ij}\}_{i<j}^{j\leq n} := \{c\mathbf{A}^{ij}\}_{i<j}^{j\leq n}$$

In the following statements, we sometimes omit the indexes of decompositions.

Lemma 21 *If* $\{\mathbf{A}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and* $\{\mathbf{B}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *are decompositions of* $\mathbf{A}$ *and* $\mathbf{B}$ *respectively, then* $\{\mathbf{A}^{ij} + \mathbf{B}^{ij}\}$ *is a decomposition of* $\mathbf{A} + \mathbf{B}$ *and* $c\{\mathbf{A}^{ij}\}$ *is a decomposition of* $c\mathbf{A}$, *for all* $c \in \mathbb{R}$.

Proof To prove the first part, one can write:

$$\sum_{i<j} [\mathbf{e}_i \;\; \mathbf{e}_j] \left(\mathbf{A}^{ij} + \mathbf{B}^{ij}\right) [\mathbf{e}_i \;\; \mathbf{e}_j]^T =$$

$$\sum_{i<j} [\mathbf{e}_i \;\; \mathbf{e}_j] \mathbf{A}^{ij} [\mathbf{e}_i \;\; \mathbf{e}_j]^T + \sum_{i<j} [\mathbf{e}_i \;\; \mathbf{e}_j] \mathbf{B}^{ij} [\mathbf{e}_i \;\; \mathbf{e}_j]^T = \mathbf{A} + \mathbf{B}$$

Moreover,

$$\sum_{i<j} [\mathbf{e}_i \;\; \mathbf{e}_j] c\mathbf{A}^{ij} [\mathbf{e}_i \;\; \mathbf{e}_j]^T = c \sum_{i<j} [\mathbf{e}_i \;\; \mathbf{e}_j] \mathbf{A}^{ij} [\mathbf{e}_i \;\; \mathbf{e}_j]^T = c\mathbf{A}$$

This proves the second part of the lemma. ■

Recall that κ is a concave function, and an analogous property of χ will be stated below.

Lemma 22 *Given the sequences* $\{\mathbf{A}^{ij}\} = \{\mathbf{A}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *and* $\{\mathbf{B}^{ij}\} = \{\mathbf{B}^{ij} \in \mathbb{S}^2\}_{i<j}^{j\leq n}$ *as well as* $c \in \mathbb{R}$, *the following properties hold:*

$$\chi(\{\mathbf{A}^{ij}\} + \{\mathbf{B}^{ij}\}) \geq \chi(\{\mathbf{A}^{ij}\}) + \chi(\{\mathbf{B}^{ij}\})$$
$$\chi(c\{\mathbf{A}^{ij}\}) \geq -|c| \max_{i'<j'} \; |\mathrm{tr}(\mathbf{A}^{i'j'})|$$

Proof Introduce

$$(i', j') = \arg\min_{i<j} \; \mathrm{tr}(\mathbf{A}^{ij});$$
$$(i'', j'') = \arg\min_{i<j} \; \mathrm{tr}(\mathbf{B}^{ij});$$
$$(i^*, j^*) = \arg\min_{i<j} \; \mathrm{tr}(\mathbf{A}^{ij} + \mathbf{B}^{ij});$$

The proof of the first inequality follows from the following expression:

$$\chi(\{\mathbf{A}^{ij}\} + \{\mathbf{B}^{ij}\}) \geq \mathrm{tr}(\mathbf{A}^{i^*j^*} + \mathbf{B}^{i^*j^*}) = \mathrm{tr}(\mathbf{A}^{i^*j^*}) + \mathrm{tr}(\mathbf{B}^{i^*j^*}) \geq$$
$$\geq \mathrm{tr}(\mathbf{A}^{i'j'}) + \mathrm{tr}(\mathbf{B}^{i''j''}) = \chi(\{\mathbf{A}^{ij}\}) + \chi(\{\mathbf{B}^{ij}\})$$

For the second inequality, one can write

$$\chi(c\{\mathbf{A}^{ij}\}) = \min_{i<j}\ \mathrm{tr}(c\mathbf{A}^{ij}) = \min_{i<j}\ c\ \mathrm{tr}(\mathbf{A}^{ij}) \geq -\max_{i<j}\ |c\ \mathrm{tr}(\mathbf{A}^{ij})| \geq -|c| \max_{i<j}\ |\mathrm{tr}(\mathbf{A}^{ij})|$$

This completes the proof. ■

Lemma 23 *If the components of the initial guess are nonzero ($\hat{x}_i \neq 0$ for all $i \in \{1,\dots,n\}$) and there exists an index $r \in \{1,\dots,m\}$ such that $\hat{\mathbf{x}}^*\mathbf{M}_r\hat{\mathbf{x}} \neq 0$, then the interior of the feasible region of Problem (8) is not empty, and strong duality holds for the penalized SOCP.*

Proof Recall that $\hat{\mathbf{x}}$ is an initial guess for the solution $\mathbf{x}$ and $\mathbf{M}$ a matrix in the objective function constructed based on $\hat{\mathbf{x}}$. We choose $c \in \{-1,+1\}$ such that $c\hat{\mathbf{x}}^T\mathbf{M}_r\hat{\mathbf{x}} > 0$, and select $\boldsymbol{\lambda} = uc\mathbf{e}_r$. It is desirable to show that if u is a sufficiently small positive number, then $\mathbf{M} + uc\mathbf{M}_r$ belongs to the interior of the $\mathcal{SDD}$ cone, i.e., it can be written as

$$\mathbf{M} + uc\mathbf{M}_r = \sum_{i<j}[\mathbf{e}_i\ \mathbf{e}_j]\mathbf{H}^{ij}[\mathbf{e}_i\ \mathbf{e}_j]^T,$$

where each $\mathbf{H}^{ij}$ is a 2×2 symmetric positive-definite matrix. By construction, the matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \sum_{i<j}[\mathbf{e}_i\ \mathbf{e}_j]\mathbf{M}^{ij}[\mathbf{e}_i\ \mathbf{e}_j]^T$$

where each $\mathbf{M}^{ij}$ is a 2×2 symmetric positive semidefinite matrix that has rank 1 and $[\hat{x}_i, \hat{x}_j]$ belongs to the null space of $\mathbf{M}^{ij}$. Now, we need to find a decomposition $\{\mathbf{B}^{ij}\}_{i<j}$ of $\mathbf{F} := c\mathbf{M}_k$ such that $\mathbf{M}^{ij} + u\mathbf{B}^{ij}$ becomes positive definite if u is small. Since the null space of $\mathbf{M}^{ij}$ is one dimensional, it suffices to show that $[\hat{x}_i\ \ \hat{x}_j]\mathbf{B}^{ij}[\hat{x}_i\ \ \hat{x}_j]^T > 0$ (due to Lemma 3.2.1 in Bertsekas (1999)). To this end, consider the following decomposition:

$$\begin{aligned}
\left[\mathbf{B}^{ij}\right]_{11} &= (d_i - d_j - F_{ij})\tfrac{\hat{x}_j}{\hat{x}_i} + \tfrac{s}{n-1}\\
\left[\mathbf{B}^{ij}\right]_{12} &= F_{ij}\\
\left[\mathbf{B}^{ij}\right]_{21} &= F_{ji}\\
\left[\mathbf{B}^{ij}\right]_{22} &= (d_j - d_i - F_{ji})\tfrac{\hat{x}_i}{\hat{x}_j} + \tfrac{s}{n-1}
\end{aligned}$$

where $d_i = \frac{\hat{\mathbf{x}}^T\mathbf{F}\mathbf{e}_i - s\hat{x}_i}{\mathbb{1}^T\hat{\mathbf{x}}}$ for every $i \in \{1,\dots,n\}$ and $s = \frac{\hat{\mathbf{x}}^T\mathbf{F}\hat{\mathbf{x}}}{\hat{\mathbf{x}}^T\hat{\mathbf{x}}}$. To complete the proof, it suffices to show that

$$\begin{cases} \mathbf{F} = \sum_{i,j}[\mathbf{e}_i\ \mathbf{e}_j]\mathbf{B}^{ij}[\mathbf{e}_i\ \mathbf{e}_j]^T \\ [\hat{x}_i\ \hat{x}_j]\mathbf{B}^{ij}[\hat{x}_i\ \hat{x}_j]^T > 0 \end{cases}$$

Which according to lemma 18 is equivalent to the following three conditions satisfied simultaneously:

$$B^{ij}_{12} = F_{ij}, \quad B^{ij}_{21} = F_{ji}, \quad \forall i<j \tag{22}$$

$$B^{ij}_{11}\hat{x}_i^2 + B^{ij}_{22}\hat{x}_j^2 > -(F_{ij}+F_{ji})\hat{x}_i\hat{x}_j \quad \forall\ i<j \tag{23}$$

$$\sum_{j<i} B^{ji}_{22} + \sum_{j>i} B^{ij}_{11} = F_{ii}, \quad \forall i \tag{24}$$

Condition (22) is straightforward to verify. To verify (23), notice that

$$\begin{aligned} B_{11}^{ij}\hat{x}_i^2 &+ B_{22}^{ij}\hat{x}_j^2 \\ &= ((d_i - d_j - F_{ij})\frac{\hat{x}_j}{\hat{x}_i} + \frac{s}{n-1})\hat{x}_i^2 + ((d_j - d_i - F_{ji})\frac{\hat{x}_i}{\hat{x}_j} + \frac{s}{n-1})\hat{x}_j^2 \\ &= (-F_{ij} - F_{ji})\hat{x}_i\hat{x}_j + s\frac{\hat{x}_i^2 + \hat{x}_j^2}{n-1} > -(F_{ij} + F_{ji})\hat{x}_i\hat{x}_j \end{aligned}$$

To analyze (24), one can write:

$$\begin{aligned} \sum_{j<i} B_{22}^{ji} &+ \sum_{j>i} B_{11}^{ij} \\ &= \sum_{j<i}((d_i - d_j - F_{ij})\tfrac{\hat{x}_j}{\hat{x}_i} + \tfrac{s}{n-1}) + \sum_{j>i}((d_i - d_j - F_{ij})\tfrac{\hat{x}_j}{\hat{x}_i} + \tfrac{s}{n-1}) \\ &= \sum_{j\neq i}(d_i - d_j - F_{ij})\tfrac{\hat{x}_j}{\hat{x}_i} + s \\ &= \tfrac{d_i}{\hat{x}_i}(\sum_j \hat{x}_j - \hat{x}_i) - \tfrac{1}{\hat{x}_i}(\sum_j d_j\hat{x}_j - d_i\hat{x}_i) - \tfrac{1}{\hat{x}_i}(\sum_j F_{ij}\hat{x}_j - F_{ii}\hat{x}_i) + s \\ &= \tfrac{d_i}{\hat{x}_i}\sum_j \hat{x}_j - \tfrac{1}{\hat{x}_i}\sum_j d_j\hat{x}_j - \tfrac{1}{\hat{x}_i}\sum_j F_{ij}\hat{x}_j + s + F_{ii} \\ &= \tfrac{d_i}{\hat{x}_i}\mathbb{1}^T\hat{\mathbf{x}} - \tfrac{1}{\hat{x}_i}\sum_j d_j\hat{x}_j - \tfrac{1}{\hat{x}_i}\mathbf{e}_i^T\mathbf{F}\hat{\mathbf{x}} + s + F_{ii} \\ &= \tfrac{d_i}{\hat{x}_i}\mathbb{1}^T\hat{\mathbf{x}} - \tfrac{1}{\hat{x}_i\mathbb{1}^T\hat{\mathbf{x}}}\sum_j[\hat{\mathbf{x}}^T\mathbf{F}\mathbf{e}_j - \hat{x}_j s]\hat{x}_j - \tfrac{1}{\hat{x}_i}\mathbf{e}_i^T\mathbf{F}\hat{\mathbf{x}} + s + F_{ii} \\ &= \tfrac{d_i}{\hat{x}_i}\mathbb{1}^T\hat{\mathbf{x}} - \tfrac{1}{\hat{x}_i\mathbb{1}^T\hat{\mathbf{x}}}[\hat{\mathbf{x}}^T\mathbf{F}\hat{\mathbf{x}} - s\hat{\mathbf{x}}^T\hat{\mathbf{x}}] - \tfrac{1}{\hat{x}_i}\mathbf{e}_i^T\mathbf{F}\hat{\mathbf{x}} + s + F_{ii} \\ &= \tfrac{d_i}{\hat{x}_i}\mathbb{1}^T\hat{\mathbf{x}} - \tfrac{1}{\hat{x}_i}\mathbf{e}_i^T\mathbf{F}\hat{\mathbf{x}} + s + F_{ii} \\ &= F_{ii}. \end{aligned}$$

As a result, if u is small, then $\|uc\mathbf{e}_r\|_\infty \leq \mu$ and $\mathbf{A}^{ij} + u\mathbf{B}^{ij}$ is positive definite. Therefore $\mathbf{M} + u\mathbf{F}$ belongs to the interior of the $\mathcal{SDD}$ cone. ■

Using the notation from Section 3.2, define

$$\mathbf{M}_r^{ij} := \begin{bmatrix} R_{ij}^r & M_{ij}^r \\ M_{ji}^r & R_{ji}^r \end{bmatrix}$$

and state the following lemma.

Lemma 24 *The sequence* $\{\mathbf{M}_r^{ij}\}_{i<j}^{j\leq n}$ *is a decomposition of* $\mathbf{M}_r$.

Proof It is straightforward to verify that

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}[\mathbf{M}_r^{ji}]_{22} + \sum_{j=2}^{n}\sum_{i=1}^{j-1}[\mathbf{M}_r^{ij}]_{11} = \sum_{j=1}^{n} R_{ij}^r = M_{ii}^r$$

The rest of the proof follows from Lemma 18. ■

Lemma 25 $\{\mathbf{M}^{ij}\}_{i<j} + \sum_{r=1}^{m}\lambda_r\{\mathbf{M}_r^{ij}\}_{i<j}$ *is a decomposition of* $\mathbf{M} + \sum_{r=1}^{m}\lambda_r\mathbf{M}_r$

Proof The proof follows immediately from Lemmas 21 and 24. ■

Proof of Lemma 6 Strong duality of the penalized SOCP follows from Lemma 23. In sight of Lemma 25, it is desirable to show that under $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$ each matrix $\mathbf{M}^{ij} + \sum \hat{\lambda}_r \mathbf{M}_r^{ij}$ is a PSD matrix. The complementary slackness condition can be written as

$$\langle [x_i \;\; x_j][x_i \;\; x_j]^T, \mathbf{M}^{ij} + \sum \lambda_r \mathbf{M}_r^{ij} \rangle = 0$$

or, given $\mathbf{M}^{ij} + \sum_{r=1}^{m} \hat{\lambda}_r \mathbf{M}_r^{ij} \succeq 0$, equivalently,

$$(\mathbf{M}^{ij} + \sum \lambda_r \mathbf{M}_r^{ij}) \begin{bmatrix} x_i \\ x_j \end{bmatrix} = \mathbf{0}. \tag{25}$$

The condition (25) combined with $\chi(\{\mathbf{M}^{ij}\}_{i<j} + \sum_{r=1}^{m} \hat{\lambda}_r \{\mathbf{M}_r^{ij}\}_{i<j}) > 0$ yields $\mathbf{M}^{ij} + \sum_{r=1}^{m} \hat{\lambda}_r \mathbf{M}_r^{ij} \succeq 0$ for all $i, j \in \{1, \ldots, n\}$, and thus $\mathbf{M} + \sum_{r=1}^{m} \hat{\lambda}_r \mathbf{M}_r \in \mathcal{SDD}$ (by Lemma 21). To satisfy the condition (25), $\boldsymbol{\lambda}$ must be such that:

$$\sum_{r=1}^{m} \lambda_r \begin{bmatrix} R_{ij}^r & M_{ij}^r \\ M_{ji}^r & R_{ji}^r \end{bmatrix} \begin{bmatrix} x_i \\ x_j \end{bmatrix} = - \begin{bmatrix} G_{ij} \\ G_{ji} \end{bmatrix} \quad \forall \; i < j$$

or equivalently

$$\sum_{r \in \mathcal{G} \cup \mathcal{B}} \lambda_r R_{ij}^r x_i = - \sum_{r \in \mathcal{G} \cup \mathcal{B}} \lambda_r M_{ij}^r x_j - G_{ij} \quad \forall \; i \neq j$$

Use the definitions given in (9) and rewrite this as

$$\tilde{\mathbf{J}}_{\mathcal{G}} \boldsymbol{\lambda}_{\mathcal{G}} = -(\tilde{\mathbf{J}}_{\mathcal{B}} \hat{\boldsymbol{\lambda}}_{\mathcal{B}} + \tilde{\mathbf{d}})$$

One solution to the above system is

$$\hat{\boldsymbol{\lambda}}_{\mathcal{G}} = -\tilde{\mathbf{J}}_{\mathcal{G}}^{+} (\tilde{\mathbf{J}}_{\mathcal{B}} \hat{\boldsymbol{\lambda}}_{\mathcal{B}} + \tilde{\mathbf{d}})$$

To conclude with dual feasibility, it is sufficient to show that

$$\chi(\{\mathbf{M}^{ij}\}_{i<j} + \sum_{r=1}^{m} \hat{\lambda}_r \{\mathbf{M}_r^{ij}\}_{i<j}) > 0,$$

which is guaranteed by condition (10b) and Lemma 22, and $\|\hat{\boldsymbol{\lambda}}\|_\infty \leq \mu$ which is guaranteed by condition (10a). On the other hand, primal feasibility is satisfied for $(\mathbf{x}\mathbf{x}^T, \eta)$. Therefore, $(\mathbf{x}\mathbf{x}^*, \eta)$ and $(\hat{\boldsymbol{\lambda}}, \{\mathbf{M}^{ij}\} + \sum_{r=1}^{m} \hat{\lambda}_r \{\mathbf{M}_r^{ij}\})$ is a primal-dual optimal pair for the problem. This completes the proof. ■

Proof of Lemma 7 Consider strong duality:

$$\langle \mathbf{x}\mathbf{x}^T, \mathbf{M}\rangle + \mu\|\boldsymbol{\eta}\|_1 = -\mathbf{y}^T\hat{\boldsymbol{\lambda}} \Longleftrightarrow$$
$$\mathbf{x}^T\mathbf{M}\mathbf{x} + \mu\|\boldsymbol{\eta}_{\mathcal{B}}\|_1 = -\sum_{r=1}^{m}\mathbf{x}^T\hat{\boldsymbol{\lambda}}_r\mathbf{M}_r\mathbf{x} - \boldsymbol{\eta}_{\mathcal{B}}^T\hat{\boldsymbol{\lambda}}_{\mathcal{B}} \Longleftrightarrow$$
$$\mathbf{x}^T\left(\mathbf{M} + \sum_{r=1}^{m}\hat{\boldsymbol{\lambda}}_r\mathbf{M}_r\right)\mathbf{x} = -\left(\mu\|\boldsymbol{\eta}_{\mathcal{B}}\|_1 + \boldsymbol{\eta}_{\mathcal{B}}^T\hat{\boldsymbol{\lambda}}_{\mathcal{B}}\right)$$

By complementary slackness condition, we have

$$\begin{aligned}\mathbf{x}^T(\mathbf{M} + \sum_{r=1}^{m}\hat{\lambda}_r\mathbf{M}_r)\mathbf{x} = & \\ &= \mathbf{x}^T\left\{\sum_{i<j}[\mathbf{e}_i \;\; \mathbf{e}_j](\mathbf{M}^{ij} + \sum\hat{\lambda}_r\mathbf{M}_r^{ij})[\mathbf{e}_i \;\; \mathbf{e}_j]^T\right\}\mathbf{x} \\ &= \sum_{i<j}[x_i \;\; x_j](\mathbf{M}^{ij} + \sum\hat{\lambda}_r\mathbf{M}_r^{ij})\begin{bmatrix} x_i \\ x_j\end{bmatrix} = 0\end{aligned}$$

Subject to the constraint $\|\hat{\boldsymbol{\lambda}}\|_\infty < \mu$, the only solution of $\mu\|\boldsymbol{\eta}_{\mathcal{B}}\|_1 + \boldsymbol{\eta}_{\mathcal{B}}^T\hat{\boldsymbol{\lambda}}_{\mathcal{B}} = 0$ is $\hat{\boldsymbol{\lambda}}_{\mathcal{B}} = -\mu\ \mathrm{sign}(\boldsymbol{\eta}_{\mathcal{B}})$ ■

Appendix C.

The next lemma will help prove some key results of the paper.

Lemma 26 *Let* $\mathbf{J}$ *be a matrix in* $\mathbb{R}^{l\times m}$, $\mathbf{d}$ *be a vector in* $\mathbb{R}^l$ *and* $\boldsymbol{\lambda}$ *be a vector in* $\mathbb{R}^m$ *such that* $\boldsymbol{\lambda}_{\mathcal{B}} = \mu\cdot\mathbf{s}$, *where* μ *is a scalar and* $\mathbf{s}$ *consists of* $+1$ *or* -1. *If*

$$\sigma_{min}(\mathbf{J}_{\mathcal{G}}) > \sigma_{max}(\mathbf{J}_{\mathcal{B}})$$

and

$$(\sigma_{min}(\mathbf{J}_{\mathcal{G}}) - \sigma_{max}(\mathbf{J}_{\mathcal{B}}))(\alpha\sigma_{min}(\mathbf{J}_{\mathcal{G}}) - \sqrt{|\mathcal{G}|}) > \sqrt{|\mathcal{B}|}\sigma_{max}(\mathbf{J}_{\mathcal{B}})\sqrt{|\mathcal{G}|} + |\mathcal{B}|\sigma_{min}(\mathbf{J}_{\mathcal{G}}) \tag{26}$$

then the interval

$$\left[\frac{\|\mathbf{d}\|_2}{\sigma_{min}(\mathbf{J}_{\mathcal{G}}) - \sigma_{max}(\mathbf{J}_{\mathcal{B}})}, \frac{(\alpha\sigma_{min}(\mathbf{J}_{\mathcal{G}}) - \sqrt{|\mathcal{G}|})\|\mathbf{d}\|_2}{\sqrt{|\mathcal{B}||\mathcal{G}|}\sigma_{max}(\mathbf{J}_{\mathcal{B}}) + |\mathcal{B}|\sigma_{min}(\mathbf{J}_{\mathcal{G}})}\right] \tag{27}$$

is not empty and the system of inequalities

$$\begin{cases} \mu > \|\boldsymbol{\lambda}_{\mathcal{G}}\|_\infty \\ \alpha\|\mathbf{d}\|_2 > \|\boldsymbol{\lambda}_{\mathcal{G}}\|_1 + \mu|\mathcal{B}| \end{cases} \tag{28}$$

is satisfied by $\boldsymbol{\lambda}_{\mathcal{G}} = -\mathbf{J}_{\mathcal{G}}^{+}(\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}} + \mathbf{b})$ *for every* μ *in the interval given in* (27).

Proof Set $\boldsymbol{\lambda}_{\mathcal{G}} = -\mathbf{J}_{\mathcal{G}}^{+}(\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}} + \mathbf{d})$ and check the set of values of μ under which the system (28) is satisfied. It can be shown that $\|\boldsymbol{\lambda}_{\mathcal{B}}\|_{\infty} = \mu$; $\|\boldsymbol{\lambda}_{\mathcal{B}}\|_2 = \mu\sqrt{|\mathcal{B}|}$. One can use several auxiliary inequalities:

1. $\|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{d}\|_1 \leq \sqrt{|\mathcal{G}|}\|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{d}\|_2 \leq \sqrt{|\mathcal{G}|}\|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{d}\|_2$
2. $\|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{d}\|_{\infty} \leq \|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{d}\|_2 \leq \|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{d}\|_2$
3. $\|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}}\|_1 \leq \sqrt{|\mathcal{G}|}\|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}}\|_2 \leq \sqrt{|\mathcal{G}|}\|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{J}_{\mathcal{B}}\lambda_{\mathcal{B}}\|_2 \leq \mu\sqrt{|\mathcal{G}||\mathcal{B}|}\|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{J}_{\mathcal{B}}\|_2$
4. $\|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}}\|_{\infty} \leq \|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{J}_{\mathcal{B}}\|_{\infty}\|\boldsymbol{\lambda}_{\mathcal{B}}\|_{\infty} \leq \mu\|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{J}_{\mathcal{B}}\|_2 \leq \mu\|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{J}_{\mathcal{B}}\|_2$

It is desirable to show that

$$\begin{cases} \mu > \|\mathbf{J}_{\mathcal{G}}^{+}(\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}} + \mathbf{d})\|_{\infty} \\ \alpha\|\mathbf{b}\|_2 > \|\mathbf{J}_{\mathcal{G}}^{+}(\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}} + \mathbf{d})\|_1 + \mu|\mathcal{B}| \end{cases} \tag{29}$$

One can use $\|\mathbf{J}_{\mathcal{G}}^{+}(\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}} + \mathbf{d})\| \leq \|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{d}\| + \|\mathbf{J}_{\mathcal{G}}^{+}\mathbf{J}_{\mathcal{B}}\boldsymbol{\lambda}_{\mathcal{B}}\|$ and relax the inequalities in (29) by applying the auxiliary inequalities:

$$\begin{cases} \mu > \|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{d}\|_2 + \mu\|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{J}_{\mathcal{B}}\|_2 \\ \alpha\|\mathbf{d}\|_2 > \sqrt{|\mathcal{G}|}\|\mathbf{J}_{\mathcal{G}}^{+}\|_2\|\mathbf{d}\|_2 + \mu\sqrt{|\mathcal{G}|}\|\mathbf{J}_{\mathcal{G}}^{+}\|_2\sqrt{|\mathcal{B}|}\|\mathbf{J}_{\mathcal{B}}\|_2 + \mu|\mathcal{B}| \end{cases}$$

Using $\|\mathbf{J}_B\|_2 = \sigma_{max}(\mathbf{J}_{\mathcal{B}})$ and $\|\mathbf{J}_{\mathcal{G}}^{+}\|_2 = \sigma_{min}(\mathbf{J}_{\mathcal{G}})^{-1}$, it yields that

$$\begin{cases} \mu(1 - \frac{\sigma_{max}(\mathbf{J}_{\mathcal{B}})}{\sigma_{min}(\mathbf{J}_{\mathcal{G}})}) > \frac{\|\mathbf{d}\|_2}{\sigma_{min}(\mathbf{J}_{\mathcal{G}})} \\ \alpha\|\mathbf{d}\|_2 > \frac{\sqrt{|\mathcal{G}|}}{\sigma_{min}(\mathbf{J}_{\mathcal{G}})}\|\mathbf{d}\|_2 + \mu\left(\frac{\sqrt{|\mathcal{G}|}}{\sigma_{min}(\mathbf{J}_{\mathcal{G}})}\sqrt{|\mathcal{B}|}\sigma_{max}(\mathbf{J}_{\mathcal{B}}) + |\mathcal{B}|\right) \end{cases}$$

One can express the bounds on μ as

$$\begin{cases} \mu > \frac{\|\mathbf{d}\|_2}{\sigma_{min}(\mathbf{J}_{\mathcal{G}}) - \sigma_{max}(\mathbf{J}_{\mathcal{B}})} \\ \mu < \frac{\alpha\sigma_{min}(\mathbf{J}_{\mathcal{G}})\|\mathbf{d}\|_2 - \sqrt{|\mathcal{G}|}\|\mathbf{b}\|_2}{\sqrt{|\mathcal{B}|}\sigma_{max}(\mathbf{J}_{\mathcal{B}})\sqrt{|\mathcal{G}|} + |\mathcal{B}|\sigma_{min}(\mathbf{J}_{\mathcal{G}})} \end{cases}$$

This gives rise to a condition to guarantee that the interval is not empty:

$$\frac{(\alpha\sigma_{min}(\mathbf{J}_{\mathcal{G}}) - \sqrt{|\mathcal{G}|})\|\mathbf{d}\|_2}{\sqrt{|\mathcal{B}|}\sigma_{max}(\mathbf{J}_{\mathcal{B}})\sqrt{|\mathcal{G}|} + |\mathcal{B}|\sigma_{min}(\mathbf{J}_{\mathcal{G}})} > \frac{\|\mathbf{d}\|_2}{\sigma_{min}(\mathbf{J}_{\mathcal{G}}) - \sigma_{max}(\mathbf{J}_{\mathcal{B}})}$$

The above inequality holds by (26). This concludes the proof. ■

Proof of Lemma 14 Note that the inequality (12) is stronger than

$$\sqrt{|\mathcal{G}|(1 - \Delta_{|\mathcal{G}|})} > \sqrt{|\mathcal{B}|(1 + \Delta_{|\mathcal{B}|})}$$

In light of Lemma 14 in Bhatia et al. (2015), any randomly sampled Gaussian matrix $\mathbf{X} \in \mathbb{R}^{l \times m}$ satisfies the inequalities

$$\lambda_{max}(\mathbf{X}\mathbf{X}^T) \leq m + (1 - 2\varepsilon)^{-1}\sqrt{cml + c'm\log\tfrac{2}{\delta}}$$
$$\lambda_{min}(\mathbf{X}\mathbf{X}^T) \geq m - (1 - 2\varepsilon)^{-1}\sqrt{cml + c'm\log\tfrac{2}{\delta}}$$

with probability at least $1-\delta$ for every $\varepsilon > 0$, where $c = 24e^2 \log \frac{3}{\varepsilon}$ and $c' = 24e^2$. This implies that the relations

$$\sigma_{min}(\mathbf{J}_{\mathcal{G}}) \in [\sqrt{|\mathcal{G}|(1-\Delta_{|\mathcal{G}|})}, \sqrt{|\mathcal{G}|(1+\Delta_{|\mathcal{G}|})}]$$

and

$$\sigma_{max}(\mathbf{J}_{\mathcal{B}}) \in [\sqrt{|\mathcal{B}|(1-\Delta_{|\mathcal{B}|})}, \sqrt{|\mathcal{B}|(1+\Delta_{|\mathcal{B}|})}]$$

are each satisfied with the probability $1-\delta$, and both are met simultaneously with probability at least $(1-\delta)^2$. By tightening the bounds in Lemma 26 with these limits on singular values, it is straightforward to verify the statement of the theorem. ■

References

Ali Abur and Antonio Gomez Exposito. *Power system state estimation: theory and implementation*. CRC press, 2004.

Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Mathematical programming*, 95(1):3–51, 2003.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2107–2116, 2017.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Jinghui Chen, Lingxiao Wang, Xiao Zhang, and Quanquan Gu. Robust wirtinger flow for phase retrieval with arbitrary corruption. *arXiv preprint arXiv:1704.06256*, 2017.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782, 2013a.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7): 4324–4337, 2013b.

Arnak Dalalyan and Yin Chen. Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems*, pages 1259–1267, 2012.

Deepjyoti Deka, Ross Baldick, and Sriram Vishwanath. Optimal data attacks on power grids: Leveraging detection & measurement jamming. In *International Conference on Smart Grid Communications (SmartGridComm)*, pages 392–397. IEEE, 2015.

Salar Fattahi and Somayeh Sojoudi. Data-driven sparse system identification. In *56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2018.

Stéphane Fliscounakis, Patrick Panciatici, Florin Capitanescu, and Louis Wehenkel. Contingency ranking with respect to overloads in very large power systems taking into account uncertainty, preventive, and corrective actions. *IEEE Transactions on Power Systems*, 28(4):4909–4917, 2013.

Mituhiro Fukuda, Masakazu Kojima, Kazuo Murota, and Kazuhide Nakata. Exploiting sparsity in semidefinite programming via matrix completion i: General framework. *SIAM Journal on Optimization*, 11(3):647–674, 2001.

Nima Hamidi and Mohsen Bayati. On low-rank trace regression under general sampling distribution. *arXiv preprint arXiv:1904.08576*, 2019.

Paul Hand and Vladislav Voroninski. Corruption robust phase retrieval via linear programming. *arXiv preprint arXiv:1612.03547*, 2016.

Roger Horn. *Matrix analysis*. Cambridge University Press, New York, 2013.

Ming Jin, Javad Lavaei, and Karl Johansson. A semidefinite programming relaxation under false data injection attacks against power grid ac state estimation. In *55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 236–243. IEEE, 2017.

Cédric Josz, Stéphane Fliscounakis, Jean Maeght, and Patrick Panciatici. AC power flow data in MATPOWER and QCQP format: iTesla, RTE snapshots, and PEGASE. *arXiv preprint arXiv:1603.01533*, 2016.

Cedric Josz, Yi Ouyang, Richard Zhang, Javad Lavaei, and Somayeh Sojoudi. A theory on the absence of spurious solutions for nonconvex and nonsmooth optimization. *Advances in Neural Information Processing Systems*, 2018.

Olga Klopp, Karim Lounici, and Alexandre B Tsybakov. Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564, 2017.

Ramtin Madani, Somayeh Sojoudi, and Javad Lavaei. Convex relaxation for optimal power flow problem: Mesh networks. *IEEE Transactions on Power Systems*, 30(1):199–211, 2014.

Ramtin Madani, Morteza Ashraphijuo, and Javad Lavaei. Promises of conic relaxation for contingency-constrained optimal power flow problem. *IEEE Transactions on Power Systems*, 31(2):1297–1307, 2016.

Ramtin Madani, Abdulrahman Kalbat, and Javad Lavaei. A low-complexity parallelizable numerical algorithm for sparse semidefinite programming. *IEEE Transactions on Control of Network Systems*, 2017a.

Ramtin Madani, Javad Lavaei, Ross Baldick, and Alper Atamtürk. Power system state estimation and bad data detection by means of conic relaxation. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017b.

Ramtin Madani, Somayeh Sojoudi, Ghazal Fazelnia, and Javad Lavaei. Finding low-rank solutions of sparse linear matrix inequalities using convex optimization. *SIAM Journal on Optimization*, 27(2):725–758, 2017c.

Ramtin Madani, Javad Lavaei, and Ross Baldick. Convexification of power flow equations in the presence of noisy measurements. *IEEE Transactions on Automatic Control*, 64(8): 3101–3116, 2019.

Brian McWilliams, Gabriel Krummenacher, Mario Lucic, and Joachim M Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423, 2014.

Hyde M Merrill and Fred C Schweppe. Bad data suppression in power system static state estimation. *IEEE Transactions on Power Apparatus and Systems*, 6:2718–2725, 1971.

Igor Molybog, Ramtin Madani, and Javad Lavaei. Conic optimization for robust quadratic regression: Deterministic bounds and statistical analysis. *IEEE 57th Conference on Decision and Control (CDC)*, 2018.

Kazuhide Nakata, Katsuki Fujisawa, Mituhiro Fukuda, Masakazu Kojima, and Kazuo Murota. Exploiting sparsity in semidefinite programming via matrix completion II: Implementation and numerical results. *Mathematical Programming*, 95(2):303–327, 2003.

Nasser M Nasrabadi, Trac D Tran, and Nam Nguyen. Robust lasso with missing and grossly corrupted observations. In *Advances in Neural Information Processing Systems*, pages 1881–1889, 2011.

Nam H Nguyen and Trac D Tran. Exact recoverability from dense corrupted observations via l_1-minimization. *IEEE transactions on information theory*, 59(4):2017–2035, 2013.

Frank Permenter and Pablo Parrilo. Partial facial reduction: simplified, equivalent SDPs via approximations of the PSD cone. *Mathematical Programming*, pages 1–54, 2014.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

Somayeh Sojoudi, Ramtin Madani, Ghazal Fazelnia, and Javad Lavaei. Graph-theoretic algorithms for polynomial optimization problems. In *IEEE 53rd Conference on Decision and Control*, pages 2257–2271. IEEE, 2014.

Christoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bolcskei. Recovery of sparsely corrupted signals. *IEEE Transactions on Information Theory*, 58(5):3115–3130, 2012.

Jan Ámos Víšek. The least trimmed squares. Part I: Consistency. *Kybernetika*, 42(1):1–36, 2006.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

Jun-Kun Wang and Shou-De Lin. Robust inverse covariance estimation under noisy measurements. In *International Conference on Machine Learning*, pages 928–936, 2014.

Yang Weng, Marija D Ilić, Qiao Li, and Rohit Negi. Convexification of bad data and topology error detection and identification problems in ac electric power systems. *IET Generation, Transmission & Distribution*, 9(16):2760–2767, 2015.

John Wright and Yi Ma. Dense error correction via l_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems*, pages 1801–1808, 2009.

Wenzhuo Yang and Huan Xu. A unified robust regression model for lasso-like algorithms. In *International Conference on Machine Learning*, pages 585–593, 2013.

Huishuai Zhang, Yuejie Chi, and Yingbin Liang. Provable non-convex phase retrieval with outliers: Median truncatedwirtinger flow. In *International conference on machine learning*, pages 1022–1031, 2016.

Richard Y Zhang, Cédric Josz, Somayeh Sojoudi, and Javad Lavaei. How much restricted isometry is needed in nonconvex matrix recovery? *Advances in Neural Information Processing Systems*, 2018a.

Yu Zhang, Ramtin Madani, and Javad Lavaei. Conic relaxations for power system state estimation with line measurements. *IEEE Transactions on Control of Network Systems*, 5(3):1193–1205, 2018b.